Towards Multilingual Vision-Language Models

Po-Yao Huang

CMU-LTI-21-011

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213
www.lit.cs.cmu.edu

Thesis Committee:

Alexander G. Hauptmann	Carnegie Mellon University
Graham Neubig	Carnegie Mellon University
Eduard Hovy	Carnegie Mellon University
Shih-Fu Chang	Columbia University

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in Language and Information Technologies

Keywords: multilingual multimodal representation, multimodal machine translation, cross-modal retrieval, adversarial learning

For my family.

Abstract

With the exploding amount of user-generated multimodal data, learning multimodal representations has enabled many novel vision-language applications in recent years. While there are around 6,500 languages worldwide, most vision-language models and their datasets are English-based. This constraint, unfortunately, hinders current models from benefiting the broader non-English community. Therefore, it is urgent yet rewarding to develop methods that generalize English-based vision-language models to non-English languages.

My thesis work makes progress on multiple fronts of this challenge via exploring the emerging trend of learning multilingual multimodal representations that facilitates modeling and reasoning over heterogeneous content including image, video, and text in various languages.

In the first part of this thesis, I identify the limitations in existing English-image representation learning to pave the path towards generalized multilingual multimodal representation learning. While prior work mainly associates whole images to the corresponding English captions, I argue such correspondence should be more fine-grained and even multilingual. The results show that learning attention-based and object-oriented multilingual multimodal representations effectively improves end tasks such as cross-modality search and multimodal machine translation.

The second part of this thesis studies cross-lingual generalizations of vision-language models. I address the scalability challenge in large-scale task-agnostic multilingual multimodal pre-training and the lack-of-annotation challenge when fine-tuning on the end task. To learn with noisy million-scale uncurated instructional videos and their transcriptions in various languages, I analyze the desirable supporting-set size in multimodal self-supervised learning and propose a reconstruction objective to alleviate such bottleneck. Additionally, I explore multilingual multimodal pre-training and construct the Multi-HowTo100M dataset, a collection of 120M video clips and their transcriptions in 9 languages, to improve zero-shot cross-lingual transfers of vision-language models. Finally, in task-specific fine-tuning, I exploit automated visual semantics to learn with sparse English-vision annotations. When non-English annotations are scarce or unavailable, I investigate visual-pivoting supervised and unsupervised multimodal machine translation to translate English-vision data into non-English-vision for multilingual multimodal fine-tuning.

The combined effort in this thesis leads to notable breakthroughs for enhancing the cross-lingual generalization capabilities of vision-language models. I believe the proposed methodologies and the resources released will be a crucial step towards multilingual vision-language models.

Acknowledgments

Many people have contributed to the work in this dissertation. I am honored to collaborate with many great minds in our time.

I would like to express my deepest appreciation to my advisor, Alex, for teaching me to bravely and creatively approach challenging yet important research problems with perseverance and an open mind. In many projects, he also set a good role model for respecting individual differences and leading the team. Alex, your wisdom, passion, and humility constantly inspire me to be a thoughtful researcher and a visionary leader.

I am grateful to my committee members: Graham, Ed, and Shih-Fu for providing me valuable insights and many helpful suggestions in shaping my dissertation. I am thankful that our paths intertwined, especially in the hardest last mile with the unprecedented COVID-19 pandemic.

Thank you to my collaborators, Xiaojun, Junjie, Mandela, Florian and many others for the fruitful discussion, timely debugging, and all the blood and tears shed before paper deadlines. I could not cross the finishing line without your vital help and encouragement.

Very special gratitude goes out to my lab mates and friends, Shoou-I, Kenneth, Lu, Danny, Junwei, Wenhe, Yijun, Lijun, Guoliang, Liangke, Ting-Yao, Salvador, Bhuwan, Mingjie, and Vaibhav for pushing projects ahead, proof-reading papers, and practicing talks. It's been an honor and a privilege working with you. I am thrilled that we achieved so many great goals together.

Finally, I would like to thank my wife, Wen-Chi Chen, and my children, Eli Huang and Eliana Huang, for your support and accompany during my Ph.D. journey. I dedicate this dissertation to you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

Learning multimodal representations is central to the success of many vision-language tasks, including cross-modal retrieval (Kiros et al., 2014; Faghri et al., 2018; Wang et al., 2016a; Vendrov et al., 2015; Nam et al., 2017; Lee et al., 2018; Huang et al., 2019d), image and video captioning (Rennie et al., 2017; Karpathy and Fei-Fei, 2015; Xu et al., 2016; Dai and Lin, 2017; Wang et al., 2019), visual question answering (Antol et al., 2015; Goyal et al., 2017; Anderson et al., 2018), multimodal semantic indexing and information extraction for multimodal content (Huang et al., 2017a; PI et al., 2018), etc. The common practice behind these vision-language models is to learn a visual-semantic embedding (VSE) space (Wang et al., 2016a) where the contents in different modalities (*e.g.*, text queries and images in cross-modal retrieval) are projected and interact with each other. In such joint embedding space, multimodal contents are mapped closely if they are semantically correlated (*e.g.*, "cat" and  . Or video and its corresponding caption(s)) and are distant if they are irrelevant (*e.g.*,  and ).

Although vision-language tasks and models have attracted much research attention in recent years, there are several challenges prevent current efforts from making an actual impact on the broader non-English communities in the real world:

(1) Existing vision-language models are English-based. While there are 6,500∼7,000 languages worldwide, most of the current multimodal datasets, such as MS-COCO (Lin et al., 2014), Flick30K (Young et al., 2014), Visual Genome (Krishna et al., 2017b), are all annotated in English. Only 7% of the world population are native English speakers. This fact inevitably hinders the applicability and generalizability of current vision-language models to the non-English

speakers, which are the majority of the world population[1]. Therefore, it is urgent yet rewarding to develop methods that could generalize these English-centered vision-language models to non-English languages.

(2) Collecting non-English multimodal data is costly. Albeit some recent works have started investigating multilingual vision-language models (Elliott et al., 2016; Wang et al., 2019), the resulting multilingual multimodal datasets are typically small (*e.g.*, 30K images and captions in 4 languages for Multi30K (Elliott et al., 2016), 20K videos and captions in 2 languages in VATEX (Wang et al., 2019)). As shown in (Huang et al., 2019c), learning robust multimodal representations relies on the availability of large-scale parallel corpora. The size and language coverage of existing multilingual multimodal datasets are therefore infeasible for developing robust multilingual vision-language models. Furthermore, since preparing visual data and hiring bilingual or multilingual annotators to construct a parallel dataset is very costly[2], multilingual multimodal data and their annotations are typically sparse and limited in size. Consequently, it will be challenging to develop annotation-efficient methods that learn from un-labeled or weakly-labeled data.

(3) Challenges in large-scale pre-training. Lastly, the exploding amount of user-generated multimodal data is left behind in current multilingual vision-language research. In the past two years, large-scale pre-training has been shown as an important step in the fields of natural language process (NLP) (Devlin et al., 2019a), computer vision (CV) (Deng et al., 2009), and image-text (Lu et al., 2019). Recently, large-scale pre-training on the large-scale video-text data (*e.g.*, 138 million video clips in the HowTo100M dataset (Miech et al., 2020)) has attracted much research attention. Unlike pre-training on clean data such as ImageNet (Deng et al., 2009) and MS-COCO (Lin et al., 2014), pre-training with user-generated videos is much more challenging as untrimmed videos and their transcriptions are inherently lengthy and noisy. Also, a robust model architecture that better models the additional temporal information in videos is still under-explored. For the textual part, likely due to its scale and the inherent noise, the multilingual user-generated and machine-translated transcriptions of large-scale video collections have never been explored until this thesis work.

[1] https://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers

[2] In our experience, the average cost of a hiring AWS worker to annotate 1 image-text pair in English is 0.5-1 USD on average. For a small dataset that contains 10K images, each with 5 captions in 2 languages, the total annotation cost is estimated to be 50-100K USD. Also, it is more expensive to hire annotators who can translate or annotate in non-English languages.

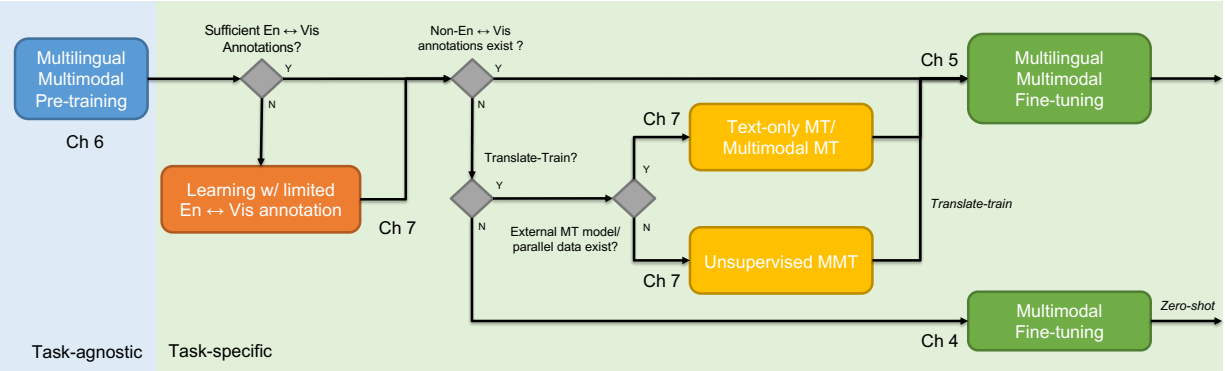


Figure 1.1: Thesis roadmap: Towards cross-lingual generalization of vision-language models.

1.2 Roadmap towards Multilingual Vision-Language Models

My work in this thesis makes progress on multiple fronts of these challenges. The goal of this thesis is to explore a practical path towards **multilingual vision-language models** which extends the scope of English-based vision-language models to cover non-English languages. As an overview, the roadmap I proposed in this thesis is shown in Fig. 1.1. To introduce multilingualism into current English-centered vision-language models, my roadmap towards multilingual vision-languages models broadly consists of two steps: (1) *task agnostic multilingual multimodal pre-training* and (2) *task-specific multilingual multimodal fine-tuning*.

In the first step, inspired by multilingual pre-training (Devlin et al., 2019a) in NLP, I study English-video and *multilingual text-video pre-training* in Chapter §6 with multilingual multimodal Transformers for learning contextual multilingual multimodal representations at a large scale. The learned representations are general-purpose and cover visual content and textual content in English and non-English languages. The Transformer encoders then can be fine-tuned for various vision-language end tasks.

In task-specific fine-tuning of multilingual vision-language models, the simplest approach is to use in-domain human-annotated non-English data when they are available. Unfortunately, in most vision-language datasets, there are no non-English annotations. To resolve this critical challenge in the path towards multilingual vision-language models, I investigate two strategies: (1) *zero-shot* and (2) *translate-train* to transfer and generalize English-centered vision-language models to non-English languages. For zero-shot cross-lingual transfer, I directly fine-tune the pre-trained multilingual multimodal Transformers on a task-specific in-domain English-vision data and apply the fine-tuned model as-is with inputs in a non-English language without any additional annotated training data in that language. On the other hand, for many languages, an

off-the-shelf machine translation (MT) or multimodal machine translation model (MMT) (*e.g.*, Chapter §7) may be available. The MT or MMT model can be used to obtain additional training data in the target language by translating the English annotations in the multimodal instances to the non-English target language. In this thesis, I am specifically interested in the multilingual setup that concerns multiple non-English languages. I fine-tune the model on the combination of original English-vision training data and all the translated non-English-vision training data. This approach is collectively categorized as the translate-train strategy discussed in the cross-lingual transfer literature in NLP (Conneau et al., 2020).

As introduced in the previous section, there are many challenges in task-specific fine-tuning. In many cases, the model must learn under limited supervision. In particular, there could be a shortage in English-vision annotations due to the cost of hiring annotators or the privacy concern of the data. Also, in most cases, there are no non-English-vision annotations. In Chapter 7, to address the sparsity issue in English-vision annotations, I introduce an annotation-efficient module to improve model robustness. Meanwhile, to enable the translate-train strategy for non-English languages, I build a supervised multimodal MT model to exploit the visual information in vision-language tasks or datasets. In the real-world scenario, off-the-shelf MT/MMT models or a parallel corpus used to train these MT models may not be available for many languages. To tackle this challenge, I further develop an unsupervised multimodal MT model to enable translate-train under that scenario.

The core of all these modules is multilingual multimodal representation learning. In the following, I briefly introduce the addressed problems in multilingual multimodal representation learning and detail the modules in the roadmap.

1.2.1 Multilingual Multimodal Representation Learning

To generalize English-based vision-language models to non-English languages, recent research efforts seek to collect parallel multilingual multimodal corpora such as Multi30K (Elliott et al., 2016) and VATEX (Wang et al., 2019). Many researchers have been investigating methods that learn cross-view representations for multiple languages. For example, Rajendran et al. (2016) learn K-view representations when parallel data is available only between one pivot view and the rest of the views. Gella et al. (2017) extend the work of (Calixto et al., 2017) to use images as the pivoting view for learning multilingual multimodal representations. Kádár et al. (2018) further confirm the benefits of multilingual training. The typical practice in prior work for learning (multilingual) multimodal representation is to respectively project paired whole image

representation to the aggregated sentence representation (*e.g.*, the average word embeddings or the last hidden state of RNN) in the shared embedding space for alignment.

In Chapter 4 and Chapter 5, I analyze the limitations of existing English-image representation learning and pave the path towards the generalized multilingual multimodal representation learning. While prior work mainly associates whole images to the corresponding English captions, I argue that such correspondence could be more fine-grained based on the observation that there are typically multiple objects in an image that are described by multiple phrases in the caption. Additionally, I design models and objectives to incorporate multilingual captions when they are available to learn multilingual multimodal representations. My results show that learning attention-based and object-oriented multilingual multimodal representations improve end tasks such as English-based (Huang et al., 2019d) text-to-image search, multilingual text-to-image search (Huang et al., 2019b).

1.2.2 Pre-training Multilingual Multimodal Representations at Scale

Existing work on multilingual multimodal representation learning such as Multi30K is inherently constrained by the dataset size. As a result, these models work only on limited languages with inferior performance compared to the English-based models. On the other hand, with the unprecedented amount of user-generated videos and captions on popular social media such as TikTok, YouTube, and Twitter, large-scale video-text pre-training (*e.g.*, HowTo100M (Miech et al., 2019)) has demonstrated great potential. Unlike image-text models, video-text models pose additional challenges. Videos are harder to models with the additional degree of freedom in time. Also, the transcriptions are noisy with the error from automatic speech recognition (ASR) and the inevitable misalignment to the visual content (Miech et al., 2020).

In chapter §6, I explore large-scale pre-training of multilingual video-text representations. I first improve contrastive learning of English-video representations by jointly reconstructing the caption from a self-excluded supporting set to alleviate the side effects in contrasting shared semantics between videos. Then I develop a task-agnostic multilingual text-video pre-training strategy by leveraging 148 million user-uploaded instructional video clips and the multilingual captions generated by users or automated MT for learning multilingual video-text representations at scale. The proposed methodology successfully promotes cross-lingual transfers of video-text models.

1.2.3 Fine-tuning under Limited Supervision

In most real-world scenarios, parallel multilingual multimodal data for a task is likely to be rare and domain-dependent. On the one hand, the amount of data available could be scarce due to privacy concerns (*e.g.*, medical data or forensic data). On the other hand, the cost to hire bilingual or multilingual professionals could be unacceptable. To this end, in Chapter 7, I develop methods to alleviate the side-effects from lack of English-vision annotations. Moreover, I leverage supervised MT and unsupervised models to translate-train and enable cross-lingual learning of vision-language end tasks.

In particular for lacking English-vision annotations, I propose to leverage the regional semantics from visual objects in an image to synthesize additional data.

When there are no non-English-vision data, I propose to translate English-vision training data into non-English languages (*i.e.*, translate-train in Fig. 1.1) via text-only MT or multimodal MT (MMT). When off-the-shelf MT/MMT models are unavailable, I further introduce an unsupervised multimodal MT module into the roadmap that relies on no parallel corpus. Specifically, I propose to exploit the multilingual image captioning and multimodal back-translation as the generators to synthesize pseudo *src-img-tgt* pairs to weakly supervise the encoder-decoder training in MMT.

In short, this thesis contributes towards enabling multilingual vision-language applications. The thesis statement is as follows: *Through learning visual-pivoting multilingual multimodal representations, we improve cross-lingual generalization capabilities of vision-language models.*

1.3 Thesis Organization

I start by introducing the background (Chapter §2) and tasks (Chapter §3) to set up the scope of this thesis. In the first part of the thesis (Chapter §4-§5), I then analyze the limitations of existing English-vision models and improve them via introducing an object-level visual prior and multilingual multimodal training. In the second part of the thesis (Chapter §6-§7), I focus on investigating effective approaches for learning multilingual video-text representations at scale (Chapter §6) and develop robust methods when a sufficient amount of annotation is unavailable (Chapter §7) The contents of each part are detailed in the following:

In Chapter §2, I first define the notations and terminologies. I then compile the literature survey for representing multimodal content (image, text, video) and present the preliminaries of cross-view contrastive learning for learning multimodal and multilingual multimodal representations. In Chapter §3, I provide an overview of popular multimodal datasets and introduce

the vision-language tasks discussed in this thesis and explain how to evaluate the task performance. Particularly, my work mainly covers cross-modal retrieval (bi-directional video-to-text and image-to-text search) and language generation tasks such as (multimodal) machine translation and captioning.

In the first part of this thesis, Chapter §4 and §5 correspond to the multilingual multimodal fine-tuning and multilingual multimodal fine-tuning in Fig. 1.1, respectively. In this part, I study (multilingual) text-image representation learning and applications on the curated multimodal datasets (Flickr30K, Multi30K, and MS-COCO). In Chapter 4, I present an adversarial probing methodology to analyze the information captured by image-text models for cross-modal retrieval. Based on the analysis, I then propose a novel object-oriented attention network (OAN) (Huang et al., 2019d) for learning visual-semantic embeddings (VSE). In Chapter 5, I demonstrate the effectiveness of learning multilingual image-text representations for the cross-modal retrieval and semantic textual similarity task. My model in (Huang et al., 2019b) features diversified multilingual VSE to improve multilingual image-text matching and grounding.

In the second part of this thesis, I address the practical challenges for building multilingual vision-language models via cross-lingual generalization of English-centered vision-language models. Specifically, I focus on: (1) multilingual multimodal pre-training at a large scale and (2) multilingual multimodal fine-tuning under limited supervision.

My work in Chapter §6 studies scalable pre-training methods on multi-million-scale user-generated videos and their captions either in English or in other languages. In section §6.2, I present my method to improve contrastive learning of English-video representations. The success of task-agnostic contrastive pre-training in NLP and CV does not apply to video-text models as videos contain rich semantics and an additional degree of freedom in time. For video-text representation learning, the supporting set to contrast across instances is sub-optimal as many semantics are shared among videos and could be ill-contrasted (*e.g.*, "man" may appear in many videos but likely to be ill-contrasted). To this end, I propose a new model in (Patrick et al., 2021b) to reconstruct captions from a self-excluded supporting-set to alleviate the side effects in video-text contrasting learning and deliver new state-of-the-art results on the retrieval and classification end tasks. In section §6.3, I then extend from the English-video scenario to the non-English-video scenario. I construct the Multi-HowTo100M dataset (Huang et al., 2021) that contains more than 100M clips and transcriptions in 9 languages. I specifically address the problem of *zero-shot cross-lingual transfer* and show that the proposed large-scale multilingual multimodal pre-training is a crucial step towards multilingual vision-language models.

Chapter §7 studies the lack-of-annotation challenges when fine-tuning multilingual multimodal

models at end tasks. To alleviate side-effects when learning with a sparsely labeled dataset where only a small portion of the dataset is with English annotations, in section §7.2, I leverage the regional visual semantics (*i.e.*, the attributes and the class names of visual objects) to generate synthetic image-text pairs for learning image-text representations. The domain gaps between original and synthesized pairs in/between modalities are closed with attention-based domain alignment with adversarial objectives (Huang et al., 2019c).

A key ingredient that enables and promotes multilingual vision-language models is machine translation (MT). In most cases, there are no non-English-vision annotations in most vision-language datasets. To enable multilingual multimodal fine-tuning with English-only annotations, in this thesis, I propose to leverage additional visual information in vision-language teaks/datasets and build up multimodal MT models to translate English training data into non-English languages for multilingual multimodal fine-tuning. For supervised multimodal MT, as one of the earliest works in this field, I validate that complementary visual information would improve text-only translation performance (Huang et al., 2016) in section §7.4. However, parallel corpus for training MT models may be unavailable for some languages. To address this issue, I introduce an unsupervised multimodal MT module in the roadmap. In section §7.4, I utilize *src-img* and *tgt-img* pairs to learn a multilingual visual-semantic embedding space for associating source and target text sub-spaces. Specifically, I explore methods of leveraging visual contents (images) as the "pseudo" pivots (Huang et al., 2020b) and learning captioning models to generate text in the source and target language for paired translation and back-translation in the absence of parallel translation annotations.

1.4 Thesis Contributions

The study in this thesis shows that, compared to English-visual representations and text-only representations, the emerging trend in learning multilingual multimodal representations has the following advantages:

Robustness Learning multilingual cross-view representations results in more generalized and robust representations. It improves multiple end tasks such as multilingual video-text (Huang et al., 2021) (Chapter §6) and image-text (Huang et al., 2019b) (Chapter §5) retrieval and multimodal MT (Huang et al., 2016, 2020b) (Chapter §7).

Grounding and Interpretability Introducing multilingualism in multimodal representation learning provides a new degree of freedom for model grounding and interpretability. For instance, (Sigurdsson et al., 2020) and my work (Huang et al., 2021) (Chapter §6) uses multilingual video-text representations to visually ground bilingual word translation. In (Huang et al., 2019b) (Chapter §5), I use multilingual image-text representations to ground multilingual text-to-image search with explicit visual object-token alignments.

Complementary Multimodal Information In multimodal MT, various works (Elliott et al., 2016; Bahdanau et al., 2015; Zhou et al., 2018c) and my work (Huang et al., 2016) (Chapter §7) shows that visual content provides complementary information that improves text-only MT.

Transferring Linguistic Knowledge via the Visual Space Even though people speak 6,500 $\sim$ 7000 languages, there is only one shared visual world since all human beings, biologically, share a similar visual system. With the availability of large-scale visual content, bridging via the visual space is a promising solution for pivoting and transferring linguistic knowledge. In this thesis I demonstrate the feasibility for (1) cross-lingual transfer of English-based vision-language models to non-English languages (Huang et al., 2021) (Chapter §6) and (2) unsupervised multimodal MT (Huang et al., 2020b) (Chapter §7).

***

In this thesis, I introduce multilingualism into current English-centered vision-language models via (1) multilingual multimodal pre-training at scale and (2) multilingual multimodal fine-tuning under limited supervision. In a nutshell, this thesis contributes towards cross-lingual generalizations of vision-language models. It provides a clear and practical roadmap with proven modules to tackle various challenges along the way. The journey begins.

Chapter 2

Background and Preliminaries

In this chapter, I introduce the background of multilingual multimodal representation learning. I organize this chapter as follows: I start with explaining the terminology and defining the notations used in this thesis. Then I introduce the typical methods for learning multimodal representations. This thesis work covers three modalities: text, image, and video. Firstly, I review the widely-used image, video, and text representations. Then I detail the existing content encoders and their architectures for each modality. Finally, I provide the formulation and the contrastive objective for learning multimodal representation (a.k.a, visual-semantic embeddings (VSE)) and introduce the multilingual scenario.

2.1 Terminologies

Prior works in different communities have very different preferences in the terminology. In this thesis, I use the term "multimodal representations" and "visual-semantic embeddings (VSE)" (where the term "semantic" here corresponds to the textual information or linguistic knowledge) interchangeably. Multimodal representations or VSE both refer to the vectorized representations of an image/video or a sentence in the visual-textual embedding space. Based on this definition, "multilingual multimodal representation" is the same as "multilingual VSE" that learns representations in a multilingual-textual-visual embedding space.

Additionally, I follow the convention in the CV community to use the term "semantics" to refer to the output of an image/video classification/detection model in Chapter §7, which could be a one-hot or multi-hot vector. Note that this is not the conventional semantics in NLP.

Also, I use the term (cross-modal) retrieval and search interchangeably to fit the terminology in the baselines I compare to. Depending on the dataset, the term "multimodal" refers to either

the image-text or video-text modality.

Moreover, I define the visual part (*e.g.*, image) and the textual part (*e.g.*, caption) as the different views of an instance. Consequently, self-supervised learning, (multimodal) contrastive learning, or (multimodal) metric learning implies similar objectives that pull paired content in different modalities closer while pushing away non-paired ones and resulting in a joint embedding space that promotes cross-modal alignments. I use contrastive learning in this thesis since I mainly study the contrastive objectives (*e.g.*, triplet ranking loss or noise contrasitve estimation (NCE)).

Lastly, like other prior works that study cross-lingual transfer in NLP, I use cross-lingual "transfer" and "generalization" interchangeably to indicate the transfer of English-vision models to the non-English-vision models. I use the term "multilingual vision-language model" to specify a vision-language model that handles various languages as inputs or outputs.

2.2 Notations

Unless otherwise specified, the notations in this thesis are defined as the following. In the English-only scenario, I use $\mathbf{w} = \{w_1, \ldots, w_N\}$ to denote the encoded word tokens where N is the sequence length. In the multilingual scenario (*e.g.*, translation tasks or multilingual text-image retrieval), I use $\mathbf{x} = \{x_1, \ldots, x_N\}$ and $\mathbf{y} = \{y_1, \ldots, y_N\}$ to denote the encoded word tokens, where $\mathbf{x}$ and $\mathbf{y}$ is the sentences in the source and target language (or English and non-English languages). For images and videos, I use $\mathbf{v} = \{v_1, \ldots, v_M\}$ as the encoded visual content, where M is the number of regions in an image or the length of a video.

For a multimodal dataset, I use $\mathcal{D} = \{\mathcal{V}, \mathcal{T}\} = \{I_1, \cdots, I_{|\mathcal{D}|}\}$ be an annotated collection of instances where each instance $I_i = (v_i, w_i)$ consists of the i-th raw image v and the corresponding raw natural language description w. For a multilingual multimodal dataset $\mathcal{D} = \{\mathcal{X}, \mathcal{Y}, \mathcal{V}\}$, the instances available are: $(\mathbf{x}, \mathbf{v}_x) \in (\mathcal{X}, \mathcal{V})$, $(\mathbf{y}, \mathbf{v}_y) \in (\mathcal{Y}, \mathcal{V})$ (drop index i for clarity), Note that for the set of images in different languages, there could be overlap under the supervised scenario $\{\mathbf{v}_x\} \cap \{\mathbf{v}_y\} \neq \phi$ where (x_i, y_i) could be translation pairs or simply captions in different languages describing the same image (Chapter §5). On the other hand, under the supervised scenario, there are no parallel translation pairs available (unsupervised), and the images are mutually exclusive for different languages, namely, $\{\mathbf{x}\} \cap \{\mathbf{y}\} = \phi$, $\{\mathbf{v}_x\} \cap \{\mathbf{v}_y\} = \phi$ (Chapter §7). The high-level aim of this thesis is to use $\mathcal{D}$ to learn multilingual multimodal representations for end tasks.

The multimodal encoders (textual encoder and visual encoder) is denoted as $\mathrm{P_T}(.)$ and $\mathrm{P_V}(.)$

Notation	Meaning	Example
w,t	raw sentence	“Three children in ...”
w_i, t_i	encoded word token	“childen”
v_j, z_j	encoded visual token	
N	max word sequence length	N=128
M	max object number or target word sequence length	M=128
K	max object number or frame length	K=128
x	raw source sentence	“Three children ...”
y	raw target sentence	“Drei Kinder ...”
v, z	raw image or video	
i	index for text	
j	index for text or visual content	
k	index for visual content	
B	mini-batch size	
$\mathcal{D}$	Multimodal dataset	
I_i	i-th image-text pair (v_i, t_i) in $\mathcal{D}$	
$\mathcal{X}$	source sentences in $\mathcal{D}$	
$\mathcal{Y}$	target sentences in $\mathcal{D}$	
$\mathcal{V}$, $\mathcal{Z}$	images or videos in $\mathcal{D}$	
$\mathbf{t}$, $\mathbf{w}$	encoded sentence	
$\mathbf{x}$	encoded source sentence	
$\mathbf{y}$	encoded target sentence	
$\mathbf{v}$, $\mathbf{z}$	encoded image or video	
$\mathbf{e}$	encoded multimodal content	
$\mathbf{h}$	hidden state	
$\mathbf{c}$	context vector	
$\mathbf{Q}, \mathbf{K}, \mathbf{V}, \mathbf{W}$	Learnable matrices in Transformer or linear layer	
w_i	attention weight	
$P_v(.)$	visual encoder or pooling function	
$P_t(.)$	textual encoder or pooling function	
$p()$	probability function	
$g()$, $h()$	inferred generator function	
$\mathcal{S}(.,.)$	similarity measure	
$\mathcal{L}(.,.)$, $l()$	loss function	
α	margin in triplet loss	
τ	softmax temperature	
λ	adjustable weights (hyper-parameter)	

Table 2.1: Notations used in this thesis.

respectively[1] to select over N encoded text and M encoded visual sequence to generate fixed-length representation for association in the contrastive learning. In contrastive learning, I use $\mathcal{L}(\mathbf{v}, \mathbf{t})$ to denote the loss operates over the text and image instances sampled in a mini-batch. I use $S(.,.)$[2] to denote the similarity measure in contrastive objective such as the triplet ranking loss or the noise contrastive estimation (NCE).

Table 2.1 summarizes the notations used in this thesis. I further use subscripts to specify the language-modality type and use the superscript to specify the objective type.

2.3 Multimodal Representation Learning

Much research has addressed learning cross-view representations for multimodal content to distill knowledge within and across modalities. Learning multimodal representation requires a mode to capture fine-grained intra-modal discrepancies and inter-modal dependencies. For instance, the system has to distinguish the phrase "*river bank*" from "*financial bank*" (intra-modal discrepancies) and connect them to the visual objects accordingly (inter-modal dependencies). I broadly divide it into two steps: Learning modality-dependent (visual and textual) representations and learning joint (visual-semantic) representations.

2.3.1 Visual and Textual Representations

Image Representation Numerous convolutional neural networks (CNN) have been explored for representing images. AlexNet (Krizhevsky et al., 2012), VGG (Simonyan and Zisserman, 2014), and ResNet (He et al., 2016b) pre-trained on ImageNet (Krizhevsky et al., 2012) are the most popular choices for extracting visual features AlexNet is used in (Yan and Mikolajczyk, 2015; Karpathy and Fei-Fei, 2015), VGG is used in (Huang et al., 2017b; Vendrov et al., 2015; Eisenschtat and Wolf, 2017) and ResNet is used in (Faghri et al., 2018; Zheng et al., 2017; Huang et al., 2017c). Instead of using global visual features, recent models (Anderson et al., 2018; Lee et al., 2018; Huang et al., 2019c,d) also explore using regional features from a Faster-RCNN (Ren et al., 2015) with a ResNet backbone pre-trained on Visual Genome (Krishna et al., 2017b). Huang et al. (2019d) show that the improvement of visual features (*i.e.*, AlexNet $\rightarrow$ VGG $\rightarrow$ ResNet) dominates the improvement of most cross-modal retrieval models.

[1] or with superscript to specify the language type in the case of multilingual training in Chapter §5

[2] I use cosine similarity $S(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{a}||\mathbf{b}|}$ is most work

Video Representation. In the video-text cross-modal retrieval, a significant change is temporal information in videos. A natural extension from 2D convolution to 3D convolution is proposed in (Ji et al., 2012) and (Tran et al., 2015) where a 3D kernel is used. Carreira and Zisserman (2017) inflate the 2D model to 3D to learn seamless spatio-temporal feature extractors from videos while leveraging successful ImageNet architecture designs and parameters. Diba et al. (2017) combine temporal information across variable depth. The R(2+1)D (Tran et al., 2018; Ghadiyaram et al., 2019) factorizes the 3D convolution filters into separate spatial and temporal components. However, the most recent works in zero-shot video retrieval still encode 2D frames individually. For example, the Dual Encoding (He et al., 2016a) uses 2D image CNN to encode frames. Chen *et al*. (Chen et al., 2020a) propose a Hierarchical Graph Reasoning (HGR) model which decomposes video-text matching into global-to-local levels but still encodes frames individually. In Chapter §6, I use a full 3D network for zero-shot video retrieval, which contains a 3D CNN-based video encoder to take care of the temporal information in videos.

Text Representation. For representing sentences, the typical approach is to use word embeddings (Mikolov et al., 2013; Pennington et al., 2014) to encode indexed word tokens followed by recurrent neural networks (RNN). Using Fisher vectors (Perronnin and Dance, 2007) of word embeddings for encoding sentences has been explored in (Frome et al., 2013). Instead of training from scratch, the 300-dimension Glove (Pennington et al., 2014) word embeddings have been used in (Zheng et al., 2017; Huang et al., 2019c) and have shown superior performance. Long-short-term memory network (LSTM) (Hochreiter and Schmidhuber, 1997) and Gated Recurrent Units (GRU) (Chung et al., 2014) are two popular RNN choices of text encoders. Note that the word embedding matrix and RNN are usually trained end-to-end.

2.3.2 Multimodal Encoders

As the number of word tokens varies sentence by sentence, a textual encoder (*i.e*., as a textual pooling function) that operates over a variable length of inputs is required for encoding sentence into a fixed-length representation for associating images or videos to the corresponding text sentence in multimodal contrastive learning to be explained in the next section. Similarly, a visual encoder (*i.e*., as a visual pooling function) is required to handle visual inputs with variable length (*e.g*., a video). Let $\{w_1, \ldots, w_N\}$ and $\{v_1, \ldots, v_M\}$ be the encoded word tokens and the encoded visual content of a image-text pair. The textual and visual pooling function ($\mathrm{P_V}(.)$ and $\mathrm{P_T}(.)$)

select over N and M respectively to generate a fixed-length representation:

$$\begin{aligned} \mathbf{v} &= \mathrm{P_V}(v_1, \ldots, v_M) \\ \mathbf{t} &= \mathrm{P_T}(w_1, \ldots, w_N) \end{aligned} \tag{2.1}$$

The simplest pooling function is *mean-pooling* or *max-pooing*. Alternatively, *attention mechanisms* provide a way to focus on certain aspects of the data per task-specific context and result in promising performance improvements. Several works (Karpathy et al., 2014a; Nam et al., 2017; Huang et al., 2017b; Lee et al., 2018) exploited different attention mechanisms for aligning visual objects and textual words. These attention mechanisms can be categorized by the type of {*Query*, *Key*, *Value* } (Vaswani et al., 2017). For intra-modal attention, {*Query*, *Key*, *Value*} are within the same modality. In DAN (Nam et al., 2017), the content in each modality is iteratively attended through multiple steps with intra-modal attention. In SCAN (Lee et al., 2018), inter-modal attention is performed between regional visual features from Anderson et al. (2018) and text semantics. The inference time complexity is $O(XY)$ (for generating X query representations for a size Y dataset). Intra-modal attention is with a preferred $O(X)$ inference time complexity compared to inter-modal attention. In this thesis, my work explore using inter-modality attention (Chapter §4), intra-modality attention (Chapter §4), as well as (intra-modal) multi-head self-attention (Chapter §5,§7) with a Transformer-based pooling (Chapter §6) to learn the textual and visual encoders.

2.3.3 Contrastive Learning

After selecting the semantically related contents from images and captions, multimodal *contrastive learning* (a type of self-supervised learning) aims at learning a shared embedding space where paired instances (*e.g.*, image and its caption) are close to each other in that space. Let $\mathcal{D} = \{I_1, \cdots, I_N\}$ be an annotated collection of instances where each instance $I_i = (\mathbf{v}_i, \mathbf{t}_i)$ consists of the image (encoded and represented as $\mathbf{v}$) and the corresponding natural language description (encoded and represented as $\mathbf{t}$). For learning a visual-semantic embedding space, most prior works including (Kiros et al., 2014; Karpathy and Fei-Fei, 2015; Wang et al., 2016b; Huang et al., 2018b) minimize a hinge-based *triplet loss* to encourage a similarity margin α between the paired visual-textual embeddings and the non-paired ones. The hinge-based triplet loss is defined as :

$$\mathcal{L}(\mathbf{v}, \mathbf{t}) = \sum_{\mathbf{v}} \sum_{\mathbf{t}^-} \left[\alpha - S(\mathbf{v}, \mathbf{t}) + S(\mathbf{v}, \mathbf{t}^-)\right]_+ + \sum_{\mathbf{t}} \sum_{\mathbf{v}^-} \left[\alpha - S(\mathbf{v}, \mathbf{t}) + S(\mathbf{v}^-, \mathbf{t})\right]_+ \tag{2.2}$$

where $[.]_+ = \max(., 0)$ is the hinge function and $S(\mathbf{v}, \mathbf{t})$ measure the similarity between $(\mathbf{v}, \mathbf{t})$ in the joint embedding space. The cosine similarity $S(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{a}||\mathbf{b}|}$ is the typical choice for the

distance/similarity measure. In Eq. 2.2, the second summations are taken over all the non-matching negative instances in another modality. This summation results in undesirable computational overhead and confuses gradient update for learning the joint embedding space.

Instead of taking summation over all the non-paired instances in the triplet loss, it is empirically shown in (Faghri et al., 2018) that taking only the hard negatives is more efficient and robust. Formally, let $(\mathbf{v}, \mathbf{t})$ denotes a paired textual-visual embeddings and $\mathbf{v}^-$ and $\mathbf{t}^-$ denote the non-matching embeddings corresponding to $(\mathbf{v}, \mathbf{t})$. The hard (non-paired) negatives can be defined as:

$$\begin{aligned} \hat{\mathbf{v}} &= \underset{\mathbf{v}-}{\text{argmax}}\ S(\mathbf{v}^-, \mathbf{t}) \\ \hat{\mathbf{t}} &= \underset{\mathbf{t}^-}{\text{argmax}}\ S(\mathbf{v}, \mathbf{t}^-) \end{aligned} \tag{2.3}$$

The triplet loss emphasizing inter-modal hard negative samples can be written as:

$$\mathcal{L}(\mathbf{v}, \mathbf{t}) = \sum_{\mathbf{v}} \big[\alpha - S(\mathbf{v}, \mathbf{t}) + S(\mathbf{v}, \hat{\mathbf{t}})\big]_+ + \sum_{\mathbf{t}} \big[\alpha - S(\mathbf{v}, \mathbf{t}) + S(\hat{\mathbf{v}}, \mathbf{t})\big]_+, \tag{2.4}$$

In practice I sample $(\mathbf{v}, \mathbf{t})$ and construct mini-batches for optimization. Formally, the objective for learning joint embeddings which promotes cross-modal alignments is defined as:

$$\mathcal{L}^{VSE}_{\mathbf{v},\mathbf{t}} = \mathbb{E}_{(\mathbf{v},\mathbf{t})\in\mathcal{D}}\Big[\mathcal{L}(\mathbf{v}, \mathbf{t})\Big] \tag{2.5}$$

At the inference time, with the trained visual encoder (generates $\mathbf{v}$) and the trained textual encoder (generates $\mathbf{t}$), the image and textural sentences are therefore be encoded for fine-tuning in other down-stream tasks or can be directly used for cross-modal retrieval such as image-text search or matching.

2.4 Multilingual Multimodal Representation Learning

Let $(\mathbf{x}^e, \mathbf{x}^g, v) \in \mathcal{D}$ be a multilingual multimodal dataset where $\mathbf{x}^e$ and $\mathbf{x}^g$ are the English and German sentence, and v is the image respectively (assume English-German-Image). Let me denote by $\mathbf{x}$ and $\mathbf{v}$ the encoded sentence and image, image-sentence similarity can be measured by the average cosine similarity $S(\mathbf{x}, \mathbf{z})$ between the visually-attend sentence embeddings and the visual embeddings. The contrastive triplet loss encouraging image-sentence alignment in the VSE space can be written as:

$$\mathcal{L}_c(\mathbf{x}, \mathbf{v}) = \max_{\tilde{\mathbf{x}}}\big[\alpha - S(\mathbf{x}, \mathbf{v}) + S(\tilde{\mathbf{x}}, \mathbf{v})\big]_+ + \max_{\tilde{\mathbf{v}}}\big[\alpha - S(\mathbf{x}, \mathbf{v}) + S(\mathbf{x}, \tilde{\mathbf{v}})\big]_+, \tag{2.6}$$

where $[.]_+$ is the hinge function, and $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{v}}$ are the non-paired (negative) instances for $\mathbf{x}$ and $\mathbf{z}$. Intuitively, when the loss decreases, the matched images and sentences will be drawn closer down to a margin α than the hardest non-paired ones. If multiple languages present, a straight forward way of learning multilingual VSE is to align English-Image $(\mathbf{x}^e, \mathbf{v}_{x^e})$ and German-Image $(\mathbf{x}^g, \mathbf{v}_{x^g})$ respectively. Formally, the following MLE objective promotes cross-modal alignments in the two VSE spaces:

$$\mathcal{L}^{VSE}_{x^e,x^g,v} = \mathbb{E}_{(\mathbf{x}^e,\mathbf{v}_{x^e})}\Big[\mathcal{L}_c(\mathbf{x}^e, \mathbf{v}_{x^e})\Big] + \mathbb{E}_{(\mathbf{x}^g,\mathbf{v}_{x^g})}\Big[\mathcal{L}_c(\mathbf{x}^g, \mathbf{v}_{x^g})\Big] \tag{2.7}$$

Note that depends on the different setup and the annotation available in a dataset, Eq. 2.7 may have a modified formulation. For example, there are two types of annotations available in Multi30K (Elliott et al., 2016).

In Chapter 5, I leverage the type II annotation in Multi30K where the descriptions in English and German are pivoted on the shared image. Therefore, I revised and Eq. 2.7 to:

$$\mathcal{L}^{VSE}_{x^e,x^g,v} = \mathbb{E}_{(\mathbf{x}^e,\mathbf{x}^g,\mathbf{v}_{x^e})}\Big[\mathcal{L}_c(\mathbf{x}^e, \mathbf{v}_{x^e}) + \mathcal{L}_c(\mathbf{x}^g, \mathbf{v}_{x^g})\Big] \tag{2.8}$$

On the other hand, for multimodal machine translation, the language part of the type I annotation is the translation pairs, which provides additional supervision signal. Intuitively, by jointly consider multimodal contrastive learning and machine translation (see Chapter §7 for more details), the objective can be rewritten as:

$$\mathcal{L}^{MMT}_{x^e,x^g,v} = \mathbb{E}_{(\mathbf{x}^e,\mathbf{x}^g,\mathbf{v}_{x^e})}\Big[\mathcal{L}_c(\mathbf{x}^e, \mathbf{v}_{x^e}) + \mathcal{L}_c(\mathbf{x}^g, \mathbf{v}_{x^g}) - \log p_{x^e \to x^g}(\mathbf{x}^g|\mathbf{x}^e) - \log p_{x^g \to x^e}(\mathbf{x}^e|\mathbf{x}^g)\Big] \tag{2.9}$$

Chapter 3

Datasets and Tasks

In this chapter, I introduce the multimodal datasets and the vision-language tasks discussed in this thesis. In section §3.1, I broadly categorize multimodal datasets into 3 categories: *Image-Text*, *Video-Text*, and *Multilingual Multimodal*. The statistics of these datasets used are compared in Table 3.1.

In section §3.2-§3.3, I cover the vision-language tasks and introduce the evaluation metrics used in this thesis. Particularly, I study cross-modal retrieval tasks and text generation tasks. In section §3.2, I introduce how to generate visual and textual representations to inference on cross-modal retrieval tasks and define the corresponding rank-based retrieval metrics. For the text generation tasks, in section §3.3, I provide the overview and comparison of objectives in supervised and unsupervised neural machine translation as well as image captioning.

3.1 Datasets

3.1.1 Image-Text Datasets

Flickr30K (Young et al., 2014) is the standard benchmark dataset for many vision-language tasks, including image search, cross-modal retrieval, image captioning, object detection, and grounding, etc. . There are 31,783 images and 158,915 image-text pairs in the Flick30K dataset. Five English descriptions are annotated for each image. The most wildly compared split in research is defined in (Karpathy and Fei-Fei, 2015) with 29,000 training, 1,000 validation, and 1,000 testing images. Flickr30k Entities (Plummer et al., 2015) additionally augment the 158k captions from Flickr30k with 244k co-reference chains, linking mentions of the same entities across different captions for the same image and associating them with 276k manually annotated bounding boxes.

Dataset	Train	Val	Test	Sent/Img	Note
			Image-Text		
Flickr30K (Young et al., 2014)	29,000	1,113	1,000	5	
MS-COCO (Lin et al., 2014)	113,287	5,000	5,000	5	
Visual Genome (Krishna et al., 2017b)	108,077	-	-	> 40	
			Video-Text		
MSR-VTT (Xu et al., 2016)	6,513	497	2,990	20	
TGIF (Li et al., 2016)	80,000	10,708	11,360	1-3	
YouCook2 (Zhou et al., 2018a)	1,333	457	210	2-16	
Youtube2text (Guadarrama et al., 2013)	1,200	100	670	5-10	
			Multilingual Multimodal		
Multi30K (Elliott et al., 2016)	29,000	1,000	1,000	5	de,fr,en,cs
VATEX (Wang et al., 2019)	25,991	3,000	6,000	10	zh,en

Table 3.1: Dataset Statistics.

MS-COCO (Lin et al., 2014) (Microsoft Common Objects in Context) is one of the most widely-used large-scale image-text datasets. It contains over 300k images and over 2.5m labeled object instances from 91 pre-defined categories. Specifically, the MS-COCO dataset contains 123,287 images where each image is annotated with five English descriptions. In total, 616,435 image-text pairs are available. The widely-compared split for image-text matching is defined by Karpathy and Fei-Fei (2015), which moves the originally left 30,504 validation images to the training set, resulting in a training set of 113,287 training images and 566,435 image-text pairs. The validation and testing split contain 5,000 images (25,000 image-text pairs) , respectively. Beyond cross-modal retrieval, MS-COCO has also been used as one of the stand benchmarks for image-captioning, object detection, object segmentation.

Visual Genome (Krishna et al., 2017b) dataset (VG) is a large scale visual dataset and consists of the components of objects, attributes, relationships, question-answer pairs. There are 108,077 images available in Visual Genome (with overlap to MS-COCO). On average, 50 regional descriptions per image (5.4 million in total). 76,340, 15,626, 47 types of object, attribute, and relationships are annotated with mapping the Wordnet Synsets. At present, VG has widely applied to scene graph generation and application for its large number of images, objects, relationships.

3.1.2 Video-Text Datasets

MSR-VTT (Xu et al., 2016) is originally developed for the video captioning task. It contains 10K generic web video clips from YouTube. MSR-VTT is collected with 257 popular video queries in 20 categories (*e.g.*sports, movie, music etc.). 200k natural sentences are describing the visual content of the clips. The average number of sentences per clip is 20, all annotated by paid human workers. In the official split, there are 6,513 clips for training, 497 clips for validation, and the remaining 2,990 clips for testing.

YouCook2 (Zhou et al., 2018a) is a collection of cooking instructional videos from YouTube. It contains 2000 long untrimmed videos from 89 cooking recipes; on average, each distinct recipe has 22 videos. The training, validation, and testing splits contain 1,333, 457, and 210 videos, respectively. The total video time is 176 hours, with an average length of 5.26 mins for each video. Each video is further trimmed into short clips as step-by-step instructions in the recipes. There are in total 14K video clips annotated with third-person view descriptions by human annotators.

HowTo100M (Miech et al., 2019) is a large-scale instructional video collection of 1.2 million Youtube videos, along with automatic speech recognition (ASR) transcriptions. There are more than 100 million clips (segmented by ASR) defined in HowTo100M. I use HowTo100M for video-text pre-training.

ActivityNet Caption (Krishna et al., 2017a) consists of densely annotated temporal segments of 20K YouTube videos. There are 10K videos for training and two validation splits that contains 5K videos each.

MSVD dataset consists of $80K$ English descriptions for 1,970 videos from YouTube, with each video associated with around 40 sentences each. I use the standard split of 1200, 100, and 670 videos for training, validation, and testing (Venugopalan et al., 2015; Xu et al., 2015; Liu et al., 2019).

3.1.3 Multilingual Multimodal Datasets

Multi30K (Elliott et al., 2016) is the multilingual extension of Flickr30K (Young et al., 2014) with additional multilingual annotations in German, French, and Czech. The training, validation, and testing split contain 29,000, 1,014, and 1,000 images, respectively. Two types of annotations are available in Multi30K: (i) One parallel English-German translation for each image. (ii) Five independently collected English and five German, French, and Czech descriptions for each image. Note that the type (ii) annotation, the descriptions are *not* translations of each other and may describe an image differently. Multi30K has been used as the standard benchmark dataset for

multimodal machine translation, multilingual image-text matching, multilingual captioning, etc. **VATEX** (Wang et al., 2019) is a collection of web videos with English and Chinese annotations. The VATEX dataset includes 25,991 videos for training, 3,000 for validation and 6,000 for testing. There are ten sentences in English and Chinese languages to describe each video. VATEX is curated for multimodal machine translation, cross-modal retrieval (video search with natural language description), and also multilingual video description generation.

3.2 Cross-Modal Retrieval Tasks

3.2.1 Text-to-image and Text-to-video Search

Cross-modal retrieval enables the search of visual content (*e.g.*, image and video) with a user-generated sentence in natural language as the query. Given a text query t, the textual encoder $\Phi(t)$ of the model encodes the query into a fixed length representation $\mathbf{t} = \Phi(t)$. The model then computes the similarity (*e.g.*, cosine similarity) between the text representation and the pre-encoded image or text $\mathbf{v} = \Psi(v)$ (where $\Psi(v)$ the visual encoder). Finally, the model generates a ranked list where the set of top-ranked images or videos are served as the final retrieval output. Note that the retrieval can be bi-direction, the query can either be natural language descriptions (*i.e.*, text-to-image retrieval) or the visual content (*i.e.*, image-to-text retrieval).

Multimodal representation learning with contrastive objectives in section §2.3 directly results in a joint visual-textual embedding space where the distance or similarity can directly be measured and compared. With the corresponding learned textual and encoder, the text query and the visual contents (image or video) can be encoded for retrieval.

3.2.2 Inference-time Complexity Analysis

It is important to keep in mind that different encoding methods will result in significantly different inference time complexity for cross-modal retrieval. For example, for searching with M text queries on size N image dataset, if all the image representations can be pre-generated and pre-stored, the inference time complexity is $O(M + N)$. In contrast, if every image representation depends on the corresponding query, then the inference time complexity will be $O(MN)$.

Recent work with inter-modal attention such as SCAN (Lee et al., 2018) and the Multimodal Transformers in VilBERT (Lu et al., 2019) fall into the first category and embed $O(MN)$ complexity which could be highly non-scalable to handle real-world query traffic. On the contrary,

models with intra-modal attention, such as DAN (Nam et al., 2017) and our works in Chapter §5-§7, fall into the latter category with $O(M + N)$ complexity which is more applicable for real-world deployment.

3.2.3 Evaluation Metrics

For cross-modal retrieval tasks, I measure rank-based performance by Recall at K (R@k), Median rank (MedR), and Mean Rank (MnR). In the case of image-text search/matching (cross-modal retrieval) on the image-text datasets, there is one and only one correct image for a text query. Given a query, recall at k (R@k) calculates the percentage of test instances for which the correct one can be found in the top-K retrieved instances. Higher R@k implies better retrieval performance. I report R@1, R@5, and R@10 following the tradition. The MedR and MnR measure the median and average rank of correct items in the retrieved ranking list, respectively, where a lower score indicates a better model. I also take the sum of all R@K as rsum to reflect the overall retrieval performance.

3.3 Language Generation Tasks

Besides the retrieval tasks, I also focus on the (sequential) classification tasks that generate natural language. Specifically, I study the case of supervised and unsupervised multimodal neural machine translation, which is a multimodal extension of convention machine translation (MT). In this section, I introduce supervised MT, unsupervised MT, and image captioning in a unified view of their mathematical formulation.

3.3.1 Neural Machine Translation

Neural machine translation (MT) with deep learning has achieved great success with the encoder-decoder architecture. The sequential nature of sentences can be effectively handled by RNN (Bahdanau et al., 2015; Sutskever et al., 2014a). Fully convolutional NMT was proposed in (Kaiser and Bengio, 2016; Kalchbrenner et al., 2016), followed by more advanced architectures such as Transformers (Vaswani et al., 2017) with self-attention and positional encoding.

Typical MT models are based on the encoder-decoder framework with attention (Bahdanau et al., 2015). Let $\mathbf{x} = (x_1, \cdots, x_N)$ denotes a source sentence and $\mathbf{y} = (y_1, \cdots, y_M)$ denotes a target sentence, where $(\mathbf{x}, \mathbf{y}) \in (\mathcal{X}, \mathcal{Y})$. The encoder-decoder model learns to estimate the

following likelihood from the source sentence to the target sentence:

$$p_{x\to y}(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{M} p(y_i|\mathbf{y}_{<i}, \mathbf{x}) \tag{3.1}$$

When a parallel corpus is available, the maximum likelihood estimation (MLE) is usually adopted to optimize the (source to target language) MT model by minimizing the following loss:

$$\mathcal{L}_{x\to y}^{MT} = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim(\mathcal{X},\mathcal{Y})} \left[-\log p_{x\to y}(\mathbf{y}|\mathbf{x})\right] \tag{3.2}$$

Among all encoder-decoder models, the Transformer (Vaswani et al., 2017) architecture recently achieves state-of-the-art translation quality. Instead of using recurrent or convolutional operations, it facilitates multi-head self-attention (Lin et al., 2017).

3.3.2 Unsupervised Machine Translation

While conventional MT systems rely on the availability of a large parallel corpus, translation with zero-resource (unsupervised MT) (Lample et al., 2018a; Artetxe et al., 2018; Lample et al., 2018b) has drawn increasing research attention. Only monolingual sentences are presented at the training and validation phase, *i.e.*, only $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in \mathcal{Y}$ are available.

Successful unsupervised MT systems share several common principles. First, they require the pre-training step to properly initialize the model and establish strong monolingual language model. For example, XLM (Lample and Alexis, 2019) utilizes the masked language model objective in BERT (Devlin et al., 2019a). MASS (Song et al., 2019) utilizes a span-based sequence-to-sequence masking objective for language model pre-training.

Second, these systems transform the unsupervised problem into a weakly or self-supervised one by automatically generating pseudo sentence pairs via back-translation (Sennrich et al., 2016a). The idea behind can be analogous to the cycle-consistency objective in CycleGAN (Zhu et al., 2017) for image-image translation with unpaired data. Specifically, let us denote by $h^*(\mathbf{y}) = (\hat{x}_1, \cdots, \hat{x}_N)$ the sentence in the source language inferred from $\mathbf{y} \in \mathcal{Y}$ such that $h^*(\mathbf{y}) = \text{argmax}\, p_{y\to x}(\mathbf{x}|\mathbf{y})$. Similarly, let us denote by $g^*(\mathbf{x}) = (\hat{y}_1, \cdots, \hat{y}_M)$ the sentence in the target language inferred from $\mathbf{x} \in \mathcal{X}$ such that $g^*(\mathbf{x}) = \text{argmax}\, p_{x\to y}(\mathbf{y}|\mathbf{x})$. Then the "pseudo" parallel sentences $(h^*(\mathbf{y}), \mathbf{y})$ and $(\mathbf{x}, g^*(\mathbf{x}))$ can be further used to train two two MT models ($\mathcal{X} \to \mathcal{Y}$ and $\mathcal{Y} \to \mathcal{X}$) by minimizing the following back-translation loss:

$$\begin{aligned}\mathcal{L}_{x\leftrightarrow y}^{BT} = &\mathbb{E}_{\mathbf{x}\sim\mathcal{X}} \left[-\log p_{y\to x}(\mathbf{x}|g^*(\mathbf{x}))\right] \\ &+ \mathbb{E}_{\mathbf{y}\sim\mathcal{Y}} \left[-\log p_{x\to y}(\mathbf{y}|h^*(\mathbf{y}))\right]\end{aligned} \tag{3.3}$$

3.3.3 Image Captioning

Image captioning models are akin to machine translation models besides the non-sequential visual encoder. Formally, an image-to-source captioning model estimates the likelihood as $p_{v\to x}(\mathbf{x}|\mathbf{v}) = \prod_{i=1}^{N} p(x_i|\mathbf{x}_{<i}, \mathbf{v})$, where $\mathbf{v}$ is the encoded image. Essentially, the captioning model learns to minimize the following loss:

$$\mathcal{L}_{v\to x}^{CAP} = \mathbb{E}_{(\mathbf{v}_x,\mathbf{x})}\left[-\log p_{v\to x}(\mathbf{x}|\mathbf{v}_x)\right] \tag{3.4}$$

3.3.4 Evaluation Metrics

Following the tradition, I adapt the widely used BLEU (Papineni et al., 2002) and METEOR (Denkowski and Lavie, 2014) as the evaluation metrics for language generation tasks. These metrics claim to align well with human judgment in language generation tasks such as translation and captioning.

BLEU (bilingual evaluation understudy) is used to evaluate the quality of target text generation (*e.g.*machine translation) with source text references. In BLEU, precision and recall are approximated by modified n-gram precision and best match length, respectively. N-gram precision measures the proportion of the matched n-grams out of the total number of n-grams in the evaluated translation then geometrically averaged. For recall, BLEU uses a brevity penalty to penalizes translations for being "too short".

METEOR (Metric for Evaluation of Translation with Explicit word Ordering) is designed to improve correlation with human judgments of machine translation quality at the segment level. METEOR evaluates a text generation model by computing a score based on explicit word-to-word matches between the generated sentence and a given reference sentence. Unlike BLEU, there are free parameters in the METEOR (namely, α, β, and γ) that can be tuned to achieve maximum correlation with human judgments.

Chapter 4

Multimodal Representation Learning

4.1 Overview

Learning multimodal representation has contributed to the success of many vision-language applications such as cross-modal search (Kiros et al., 2014; Faghri et al., 2018; Lee et al., 2018; Huang et al., 2018b, 2019d), image and video captioning (Rennie et al., 2017; Karpathy and Fei-Fei, 2015; Dai and Lin, 2017), visual question answering (Goyal et al., 2017; Anderson et al., 2018), multimodal semantic indexing (Li et al., 2021; Liu et al., 2020), and information extraction for multimodal content (Huang et al., 2017a; Liang et al., 2017; PI et al., 2018; Huang et al., 2018a). The embedding model aims at encoding and fusing knowledge of multimodal contents (*e.g.*, image, video, and text) into a shared embedding space (*i.e.*, visual-semantic embedding (VSE) space) in which the distance between instances can be measured and compared. Ideally, the semantically close elements will be embedded closely to each other. To learn VSEs, the transformation function is typically trained by aligning paired multimodal inputs (*e.g.*, images and text) in a latent space where the embeddings are close if they are semantically associated or distant if uncorrelated[1].

Although significant progress has been made in image-text representation learning in recent years, few have explored what those models truly learn and what makes one model superior to another. In this chapter, I showcase my analysis methodology via adversarial probing image-text models and present an improved object-oriented image-text representation learning. I organize this chapter as follows: In Section 4.2, I present an analysis framework by probing the model trained without a specific type of textual information. In Section 4.3, based on the analysis, I

[1]Readers may refer Section 2.3 for more details about prior work.

present an object-oriented network to associate the textual and visual content for joint embedding space learning in which I achieved state-of-the-art results at that time.

The content in this chapter appears in:

1. Improving What Cross-Modal Retrieval Models Learn through Object-Oriented Inter- and Intra-Modal Attention Networks," Po-Yao Huang, Vaibhav, Xiaojun Chang, Alexander Hauptmann, *ACM ICMR 2018*

4.2 Adversarial Probing of Image-Text Models

4.2.1 What Do Image-Text Models Learn?

In the past few years, a rich line of research including advanced encoders (Zheng et al., 2017; Lee et al., 2018), attention mechanisms (Nam et al., 2017; Huang et al., 2017c, 2019d), and structure-preserving loss functions (Klein et al., 2015; Wang et al., 2016b, 2018b) or multi-tasking (Huang et al., 2018a) promoted sophisticated approaches to learn improved visual-semantic embeddings. However, few have systematically explored how those models learn the joint embedding space and what aspects of data and model are critical. I start by training two state-of-the-art text-to-image retrieval models with adversarial text inputs, I investigate and quantify the importance of syntactic structure and lexical information in learning the joint visual-semantic embedding space for cross-modal retrieval.

To shed light on the black boxes of cross-modal retrieval models and improve them, I present analyses through adversarial probing inspired by (Jia and Liang, 2017; Cirik et al., 2018). Different from these works, I jointly leverage adversarial inputs at the training phase and the leave-one-out technique for cross-validation to probe and quantify the importance of different types of information. Mostly, I are interested in how the state-of-the-art models utilize textual information to correlate the visual and textual context. In practice, I conduct experiments where various aspects of the input sentences are perturbed for training and measure the performance drop in comparison to the original model to quantify the importance of syntactic structure and lexical information.

4.2.2 Adversarial Perturbation

In the perturbation experiments, I take the publicly available codes and the features from the two representative state-of-the-art models: SCAN (Lee et al., 2018) and VSE++ (Faghri et al., 2018).

I train them from scratch with the best settings available. Both of the models utilize RNN as the text encoders in the hope that dependencies among the word tokens in a sentence could be captured and encoded accordingly. SCAN performs fine-grained alignments between encoded word tokens and visual objects. In contrast, VSE++ utilizes the last hidden state of RNN as the sentence embeddings. For visual embeddings, SCAN uses pre-extracted regional visual features from a Faster RCNN in (Anderson et al., 2018) while VSE++ fine-tunes a ResNet (He et al., 2016b).

I examine the models with the standard training, validation, and testing splits in Flickr30K (Young et al., 2014). The models are evaluated using Recall at k (R@k) metric in the text-to-image retrieval task. In the training phase, I deliberately block the model from learning a specific type of information in the data. In each run, I apply one type of perturbation to all the training sentences. The perturbation is either random shuffling of word order or dropping one type of part-of-speech (POS) tagging in the sentences. For example, I trained the model with the sentence "*shorts in dressed a baby blue ...*" which is randomly shuffled from "*a baby dressed in blue shorts ...*". In the validation and testing phase, the models are evaluated with original non-perturbed sentences. As a consequence, the model trained with perturbed inputs will perform worse than the original model. The performance degradation then provides informative clues for the importance of the type of information dropped. A significant drop in performance indicates that the blocked information is critical, whereas a small drop implies irrelevance.

VSE++ (Faghri et al., 2018)	R@1 (Δ, Δ%)	R@5 (Δ, Δ%)	R@10 (Δ, Δ%)
Original model	39.3 (0.0, -0.0%)	68.8 (0.0, -0.0%)	79.0 (0.0, -0.0%)
Random shuffling	37.3 (-2.0, -5.0%)	65.6 (-3.2, -4.6%)	76.5 (-2.5, -3.2%)
Drop Numbers	38.5 (-0.8, -2.1%)	67.4 (-1.4, -2.0%)	77.5 (-1.5, -1.9%)
Drop Prepositions	37.5 (-1.8, -4.6%)	67.0 (-1.8, -2.6%)	76.8 (-2.2, -2.8%)
Drop Verbs	36.2 (-3.1, -7.9%)	65.5 (-3.3, -4.8%)	76.2 (-2.6, -3.3%)
Drop Adjectives	34.0 (-5.3, -13.5%)	63.3 (-5.5, -8.0%)	72.4 (-6.6, -8.4%)
Drop Nouns	9.5 (-29.8, -76%)	25.7 (-43.1, -63%)	34.5 (-44.5, -56%)

Table 4.1: Performance comparison of VSE++ (Faghri et al., 2018) trained with perturbed inputs in the text-to-image retrieval task on Flickr30K. To understand what do the models learn, in each experiment, one type of perturbation is applied to the training textual inputs. Validation and testing inputs are without perturbation.

SCAN (Lee et al., 2018)	R@1 (Δ, $\Delta\%$)	R@5 (Δ, $\Delta\%$)	R@10 (Δ, $\Delta\%$)
Original model	44.6 (0.0, -0.0%)	74.1 (0.0, -0.0%)	82.5 (0.0, -0.0%)
Random shuffling	42.5 (-2.1, -4.7%)	71.2 (-2.9, -3.9%)	81.3 (-1.2, -1.5%)
Drop Numbers	43.9 (-0.7, -1.6%)	73.5 (-1.0, -1.3%)	81.7 (-0.8, -1.0%)
Drop Prepositions	42.2 (-2.3, -5.2%)	71.0 (-3.1, -4.2%)	80.0 (-2.5, -3.0%)
Drop Verbs	41.3 (-3.3, -7.4%)	70.3 (-3.8, -5.1%)	79.5 (-3.0, -3.6%)
Drop Adjectives	38.6 (-6.0, -13.5%)	67.5 (-6.6, -8.9%)	75.8 (-6.7, -8.1%)
Drop Nouns	9.8 (-34.8, -78%)	28.5 (-45.6, -62%)	37.1 (-45.4, -55%)

Table 4.2: Performance comparison of SCAN (Lee et al., 2018) trained with perturbed inputs in the text-to-image retrieval task on Flickr30K.

4.2.3 Importance of Syntactic Structure

Word order and syntactic structure in English is vital for humans to understand the meaning of a sentence correctly. Previous studies have demonstrated that encoders with recurrent architectures can capture syntactic structures and perform well in multiple text-only tasks such as named entity recognition (Lample et al., 2016), sentiment analysis (Zhang et al., 2018b), and machine translation (Bahdanau et al., 2015). In this study, I investigate whether encoding syntactic structures contribute to learning better visual-semantic embeddings for cross-modal retrieval.

Table 4.1 and Table 4.2 summarizes the impacts of perturbation for the two models of concern. Comparing the performance drop (Δ) and its relative percentage ($\Delta\%$) between the baseline and the model trained with randomly shuffled inputs, the drops in all metrics are surprisingly little ($\leq 5\%$ for R@1) for both of the models. The results show that purging the syntactic information does not bring severe degradation. In other words, even with RNNs, these models only minimally utilize syntactic information in learning the joint embedding space for aligning sentences and images.

4.2.4 Importance of Lexical Information

What do cross-modal retrieval models learn if the syntactic structure is unimportant for learning the joint embedding space? I suspect that the insignificance of syntax may also imply that spatial relationships (*e.g.*, *on*, *in*, *at*) which are usually described by prepositional words (*IN*) in the sentences may also be irrelevant. For verification, I conduct the leave-one-POS-out experiments to quantify the importance of different types of lexical information. The prior probabilities for a

lexical type to appear in a sentence are: Cardinal Numbers:21.3%, Prepositions: 92.8%, Verbs: 91.5%, Adjectives/Adverbs: 86.6%, Nouns: 99.9% in Flickr30K. For a fair comparison, the prior probabilities of different lexical types in Flickr30K are summarized in Table 4.3.

Lexical category	Counts	In # of sentences	Prior (%)
Count	39,611	33,873	21.3
Prep	225,133	147,513	92.8
Verb	284,669	145,347	91.5
Adj	295,647	137,660	86.6
Noun	650,207	158,904	99.9

Table 4.3: Statistics of lexical categories in Flickr30K

From Table 4.1 and Table 4.2, I observe vast differences among the performance drops between the original model and the model trained with inputs discarding specific types of POS tagging. Although the prior probability is low, cardinal numbers (*CD*) may not be learned by the models. Unsurprisingly, even with a higher prior probability, prepositions (*IN*) and spatial relationships in the sentences are ignored by the model. Notably, the verbal information is also not useful. The models do not fully capture the described action.

In contrast, I observe that adjectives and nouns play essential roles in cross-modal retrieval. Even with a relatively smaller prior, dropping adjectives results in more than 13% performance drop in R@1 for both of the models. As expected, the most critical lexical type for cross-modal retrieval is noun.

Based on the significance of adjectives and nouns, I conclude that those models mainly learn to leverage noun phrases to bridge across modalities in the joint embedding space. A bridge connects two shores. At the textual side, noun phrases are the proper units for encoding and linking. At the visual side, as simple descriptors (*e.g.*, red) and visible objects (*e.g.*, apple) are most often described by adjectives and nouns; visual objects are likely to be the proper units. The superiority of SCAN (Lee et al., 2018) over (Faghri et al., 2018) may result from the regional object-level visual features extracted by Faster RCNN (Ren et al., 2015). Driven by this observation, I argue that the alignments in the joint embedding space are likely to be object-oriented. I expect a model promoting object-oriented links would help to learn a better joint embedding space for cross-modal retrieval.

4.3 Learning Object-Oriented Image-Text Representations

The results in Section 4.2 show that the retrieval power mainly comes from localizing and connecting the visual objects and their cross-modal counter-parts, the textual phrases. Inspired by this observation, I propose a novel model that employs object-oriented encoders along with inter- and intra-modal attention networks to improve inter-modal dependencies for cross-modal retrieval. Besides, I develop a new multimodal structure-preserving objective, which additionally emphasizes intra-modal hard negative examples to promote intra-modal discrepancies.

As demonstrated in the perturbation experiments, utilizing object-oriented textual and visual context is crucial for cross-modal retrieval. Following this insight, I propose to jointly use CNN for encoding textual context and Faster RCNN for visual context. I propose to use inter-modal attention to refine, compare, and align these regionally-encoded objects. On the other hand, the global context, which encodes interactions among multiple regional objects, is not well-captured by RNN in (Faghri et al., 2018) or CNN with max-pooling in (Ma et al., 2015; Zheng et al., 2017). To address this issue, I propose to employ a multi-head context-aware intra-modal attention network to distill the crucial global context for alignment. I show that incorporating both regional and global context with inter- and intra-modal attention will learn better visual-semantic embeddings. Additionally, I propose a new triplet objective that emphasizes inter and intra-modal harder negatives to retain both the cross-view and the within-view structure in the joint embedding space. I term the proposed model object-oriented attention network (**OAN**). The overall architecture is illustrated in Fig. 4.1.

4.3.1 Object-Oriented Encoders

For representing textual objects (*i.e.*, noun phrases), instead of using RNNs or recursive neural network (Niu et al., 2017) to traverse the parsing tree to encode word sequences. However, I propose to use simple CNN-based encoders that better captures phrases. Given a sentence $\mathcal{S} = [t_1 ... t_N]$, I encode the word through an embedding matrix $x_i = \mathbf{W}_W t_i, i \in \{1...N\}$. Then for each time step, I feed the word embeddings in the sentence to a set of 1-D convolution kernels $\{\mathbf{K}_j\}, j \in \{1, 2, 3, 5, 7\}$. Formally, I have:

$$\mathbf{H}^j = \tanh(\mathbf{K}_j * \mathbf{X} + b_j) \tag{4.1}$$

where "*" is the convolution operation and b_j is the bias. The j-th convolutional kernels aim to capture features corresponding to uni-grams ($j = 1$), bi-grams ($j = 2$) and so on. In each time step i, I concatenate the outputs h_i^j into $h_i = h_i^1 \| h_i^2 \| h_i^3 \| h_i^5 \| h_i^7$ and transform it to the textual

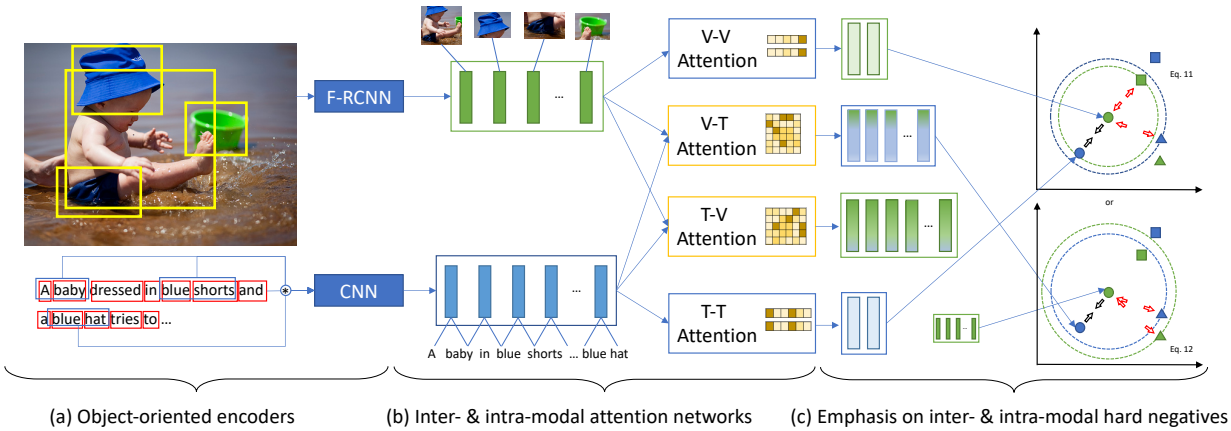


Figure 4.1: The proposed object-oriented attention network (**OAN**) is composed of 3 components: (a) Object-oriented encoders (§ 4.3.1). (b) Inter- and intra-modal attention networks (§ 4.3.2). (c) Inter- and intra-modal hard negative mining (§ 4.3.3). Convolution kernels for uni-gram and bi-gram are colored in red and blue. Visual and textual embeddings are colored in green and blue, respectively. Attention weights are proportional to the darkness in yellow. Different shapes in (c) indicate different instances. (Better viewed in color.)

embedding $e_i \in \mathbb{R}^D$ via a single-layer perceptron parameterized by $\mathbf{W}_T$ and b_T. The i-th textual object embedding e_i is:

$$e_i = \tanh(\mathbf{W}_T h_i + b_T) \tag{4.2}$$

For representing visual objects, I follow the practice in (Anderson et al., 2018) to localize salient objects with Faster RCNN (Ren et al., 2015) and extract regional features with a ResNet-101 (He et al., 2016b) backbone. For an image $\mathcal{U}$, I use the pre-trained model provided by (Anderson et al., 2018) to detect M visual objects with top confidence scores and get the regional features for each object. With $\mathcal{U} = \{u_1 ... u_M\}, u_j \in \mathcal{R}^{D_V}$ from Faster RCNN, I then apply a single-layer perceptron parameterized by $\mathbf{W}_V$ and b_V to map them into D-dimensional embeddings. The j-th visual object embedding v_j is:

$$v_j = \tanh(\mathbf{W}_V u_j + b_V) \tag{4.3}$$

4.3.2 Inter-Modal and Intra-Modal Attention

Inter-modal attention is useful to build regional connections between textual and visual objects. However, those inter-modal connections are not equally important, and the interactions between objects within the same modality are missing. Intra-modal attention, as firstly introduced by (Nam

et al., 2017) in the context of cross-modal retrieval with a multi-hop memory network, serves as a pooling function to pack the regional context with a global view. I propose to jointly leverage these two types of attention to better aligning objects in the two modalities.

Inter-modal attention: There are two directions in inter-modal attention: textual-to-visual (T-V) and visual-to-textual (V-T) attention. The goal of applying inter-modal attention is to transform the embedding of an object in one modality according to its relevance to the objects from the other modality. By minimizing the distance between the attended embedding and the original embedding, the model achieves fine-grained alignments between objects in different modalities in the joint embedding space. The attention networks take the outputs from the object-oriented encoders: $\mathcal{E} = \{e_1, ..., e_N\}, e_i \in \mathbb{R}^D$ and $\mathcal{V} = \{v_1, ..., v_M\}, v_j \in \mathbb{R}^D$. Let $s_{ij} = \frac{[e_i^T v_j]_+}{\|e_i\|\|v_j\|}$ denotes the similarity between the i-th textual phrase and the j-th visual object, where $i \in \{1, ..., N\}, j \in \{1, ..., M\}$ and $[.]_+ = max(., 0)$ is the hinge function.

In V-T attention, visually-attended textual embeddings e^V are generated as a weighted combination of textual embeddings. The visually-attended weights are calculated by the softmax function with a temperature parameter λ. Specifically,

$$w_{ij} = \text{softmax}(\lambda \hat{s}_{ij}), \tag{4.4}$$

$$e_j^V = \sum_{i=1}^{N} w_{ij} e_i, j \in \{1...M\}, \tag{4.5}$$

where $\hat{s}_{ij} = s_{ij}/\|s_{i,:}\|_2$. Similarly, for T-V attention, I generate the textually-attended visual embeddings v_i^T as a weighted combination of visual embeddings with $\bar{s}_{ij} = s_{ij}/\|s_{:,j}\|_2$.

There are two sets of object-wise embeddings in each direction of inter-modal attention to represent an instance. For the V-T attention network, the instance-wise visual and textual embeddings are $v_{VT} = \{v_1, ..., v_M\}$ and $e_{VT} = \{e_1^V, ..., e_M^V\}$. For the T-V attention network, the instance-wise visual and textual embeddings are $v_{TV} = \{v_1^T, ..., v_N^T\}$ and $e_{TV} = \{e_1, ..., e_N\}$, respectively.

Intra-modal attention: The intra-modal attention network focuses on certain aspects of data concerning the intra-modal context. In contrast to (Nam et al., 2017), I leverage a single-hop multi-head context-aware attention network to capture the interactions between objects with a global view and to distill informative objects for alignment from individual modalities. Let $c_T = \frac{1}{N}\sum_i e_i$ and $c_V = \frac{1}{M}\sum_j v_j$ denote the the textual and visual context respectively. For k-th head of textual-to-textual attention (T-T), I define the k-th textually attended textual embedding as a weighted combination of textual embeddings:

$$h_i = \tanh(\mathbf{W}_{ct}^k c_T)^T \tanh(\mathbf{W}_t^k e_i), \tag{4.6}$$

$$w_i^k = \text{softmax}(\lambda h_i), \tag{4.7}$$

$$e_{TT}^k = \sum_{i=1}^{N} w_i^k e_i \tag{4.8}$$

Similarly, I generate k-th visually attended visual embedding v_{VV}^k with visual-to-visual attention (V-V). The final instance-wise intra-modal attended representations are: $e_{TT} = \{e_{TT}^1, ..., e_{TT}^K\}$ and $v_{VV} = \{v_{VV}^1, ..., v_{VV}^K\}$ for the textual and visual part, respectively.

4.3.3 Emphasis on Inter-Modal and Intra-Modal Hard Negatives

To further promote intra-modal discrepancies, I propose to incorporate both the inter-model hard negative examples and the intra-model hard negative examples. Essentially, given t, I sample the intra-modal hard negative example $\bar{t}$ paired to either (1) the original text embedding t or (2) the text part of the hard visual embedding $\hat{v}$. These two types of intra-modal hard negative examples are illustrated in Fig. 4.1-(c). Let $\bar{t}$ and $\bar{v}$ denote the visual and textual intra-modal hard negative examples, I introduce two additional triplet losses $\alpha - S(v,v) - S(v,\bar{v})$ and $\alpha - S(t,t) - S(t,\bar{t})$ weighted by a hyper parameter β. Since $S(t,t) = S(v,v) = 1$. I propose the following loss function:

$$\begin{aligned} l(v,t) = &\sum_v \left\{ \left[\alpha_{inter} - S(v,t) + S(v,\hat{t})\right]_+ + \beta\left[\alpha_{intra} - S(v,\bar{v})\right]_+ \right\} \\ &+ \sum_t \left\{ \left[\alpha_{inter} - S(v,t) + S(\hat{v},t)\right]_+ + \beta\left[\alpha_{intra} - S(t,\bar{t})\right]_+ \right\} \end{aligned} \tag{4.9}$$

As depicted in Fig. 4.1-(c), I sample $\bar{t}$ and $\bar{v}$ as either the hard negative to the original textual and visual embedding:

$$\bar{v} = \underset{v^-}{\text{argmax}}\ S(v^-, v), \bar{t} = \underset{t^-}{\text{argmax}}\ S(t^-, t), \tag{4.10}$$

or the corresponding visual part of the hard negative textual embeddings and the corresponding textual part of the hard negative visual embeddings:

$$\begin{aligned} \bar{v} = v_{\hat{t}}, \hat{t} = \underset{t^-}{\text{argmax}}\ S(v, t^-), \\ \bar{t} = t_{\hat{v}}, \hat{v} = \underset{v^-}{\text{argmax}}\ S(v^-, t) \end{aligned} \tag{4.11}$$

	R@1	R@5	R@10
Full model (Eq. 4.9 and Eq. 4.10)	53.1	79.9	86.8
Full model (Eq. 4.9 and Eq. 4.11)	**53.3**	**80.1**	**87.1**
No intra-modal hard negatives	52.8	79.6	86.8
No inter-modal attention	48.2	77.5	85.5
No intra-modal attention	47.8	77.0	85.2
No attention	37.2	70.7	76.2
No object-oriented visual encoder	41.2	70.9	80.0
No object-oriented textual encoder	51.0	78.5	85.2

Table 4.4: Ablation studies of the proposed model for text-to-image retrieval in the 1K testing set of Flickr30K.

4.3.4 Empirical Evaluation

Ablation Studies I perform ablation studies in the text-to-image retrieval task on the Flickr30K testing set to quantify the contribution of each component, Each time I remove one component and measure its relative performance drop. More critical the removed component is if there is a more substantial degradation. On the top two rows, I list the full model with the two sampling strategies for intra-modal hard negative examples in the proposed objective. For the experiment without object-oriented textual encoder, I swap text CNN to Bi-directional LSTM. For the experiment without object-oriented visual encoder, I follow Nam et al. (2017) to resize images to 448×448 and use ResNet-152 to extract $14 \times 14 \times 2048$ grid-based visual features. In the experiments without attention, I employ mean pooling over the encoded objects in individual modalities directly. Table 4.4 summarizes the result.

1. **Object-oriented encoders.** Textual phrases and visual objects are central to visual-semantic embeddings. The CNN architecture in the textual encoder captures relevant linguistic priors such as n-gram structures and noun phrases. On the other hand, the object detection network (Faster RCNN), which can be analogized to the inherited natural attention in the human's cognitive system, helps the model to focus on the salient visual objects. Empirically the visual encoder plays a more critical role than the textual encoder. Both of them are crucial to achieving the best performance.
2. **Inter- and intra-modal attention mechanisms.** Instead of bluntly aligning all the context in the joint embedding space, the attention mechanisms serve as the fundamental approach

to pool and connect encoded objects. The model without attention results in the worst performance in the ablation experiments. I note that models with inter-modal and intra-modal attention yield comparable performance. The two types of attention together contribute to the top-performing model in a complementary manner.

3. **Intra-modal hard negative examples.** Preservation of the intra-modal structure achieves consistent gain, but it is not as significant as other components. The training objective considering the paired intra-modal content of the sampled inter-modal hard negative (Eq. 4.11) results in slightly higher performance than the objective considering the hard intra-modal content (Eq. 4.10).

	Text-to-Image			Image-to-Text		
	R@1	R@5	R@10	R@1	R@5	R@10
DCCA (Yan and Mikolajczyk, 2015) (TF-IDF, AlexNet)	12.6	31.0	43.0	16.7	39.3	52.9
DVSA (Karpathy and Fei-Fei, 2015) (RNN, AlexNet)	15.2	37.7	50.5	22.2	48.2	61.4
SM-LSTM (Huang et al., 2017b) (RNN, VGG)	30.2	60.4	72.3	42.5	71.9	81.5
2WayNet (Eisenschtat and Wolf, 2017) (GMM, VGG)	36.0	55.6	-	49.8	67.5	-
VSE++ (Faghri et al., 2018) (RNN, ResNet)	39.6	-	79.5	52.9	-	87.2
DPC (Zheng et al., 2017) (CNN, ResNet)	39.1	69.2	80.9	55.6	81.9	89.0
DAN (Nam et al., 2017) (RNN, ResNet)	39.4	69.2	79.1	55.0	81.8	89.5
SCO (Huang et al., 2017c) (RNN, ResNet)	41.1	70.5	80.1	55.5	82.0	89.3
SCAN (Lee et al., 2018) (RNN, FRCNN-ResNet)	45.8	74.4	83.0	61.8	87.5	93.7
OAN (Ours) (CNN, FRCNN-ResNet)	**53.3**	**80.1**	**87.1**	**68.6**	**93.0**	**96.0**

Table 4.5: Performance comparison on Flickr30K's 1K testing set. For each baseline, the best single model with highest R@1 in text-to-image retrieval task is reported and compared. I also list the backbone encoders (textual encoder, visual encoder).

Main Results

I compare various research from classic models (Yan and Mikolajczyk, 2015; Karpathy and Fei-Fei, 2015) to recent models with advanced visual features (Wang et al., 2016b, 2018b; Faghri et al., 2018; Huang et al., 2017b) and attention mechanisms (Nam et al., 2017; Huang et al., 2017c; Gu et al., 2018; Lee et al., 2018). For reference, I also list their corresponding textual and visual encoders. I either directly report the scores of the best single model in the original paper or refer to the summarization made in (Zheng et al., 2017). Note that ensembled models reported in

	Text-to-Image			Image-to-Text		
	R@1	R@5	R@10	R@1	R@5	R@10
Model	1K Testing Images					
DVSA (Karpathy and Fei-Fei, 2015) (RNN, AlexNet)	27.4	60.2	74.8	38.4	69.9	80.5
SM-LSTM (Huang et al., 2017b) (RNN, VGG)	30.2	60.4	72.3	42.5	71.9	81.5
Order-embeddings (Vendrov et al., 2015) (RNN, VGG)	37.9	-	85.9	46.7	-	88.9
2WayNet (Eisenschtat and Wolf, 2017) (GMM, VGG)	39.7	63.3	-	55.8	75.2	-
VSE++ (Faghri et al., 2018) (RNN, ResNet)	52.0	-	92.0	64.6	-	95.7
DPC (Zheng et al., 2017) (CNN, ResNet)	47.1	79.9	90.0	65.6	89.8	95.5
GXN (Gu et al., 2018) (RNN, ResNet)	56.6	-	94.5	68.5	-	97.9
SCO (Huang et al., 2017c) (RNN, ResNet)	56.7	87.5	**94.8**	69.9	92.9	97.5
SCAN (Lee et al., 2018) (RNN, FRCNN-ResNet)	56.4	87.0	93.9	70.9	94.5	97.8
OAN (Ours) (CNN, FRCNN-ResNet)	**60.2**	**88.6**	94.5	**71.7**	**96.4**	**99.3**
Model	5K Testing Images					
DVSA (Karpathy and Fei-Fei, 2015) (RNN, AlexNet)	10.7	29.6	42.2	16.5	39.2	52.0
Order-embeddings (Vendrov et al., 2015) (RNN, VGG)	31.7	-	74.6	23.3	-	84.7
VSE++ (Faghri et al., 2018) (RNN, ResNet)	30.3	-	72.4	41.3	-	81.2
DPC (Zheng et al., 2017) (CNN, ResNet)	25.3	53.4	66.4	41.2	70.5	81.1
GXN (Gu et al., 2018) (RNN, ResNet)	31.7	-	74.6	42.0	-	84.7
SCO (Huang et al., 2017c) (RNN, ResNet)	33.1	62.9	75.5	42.8	72.3	83.0
SCAN (RNN, FRCNN-ResNet)	34.4	63.7	75.7	46.4	77.4	87.2
OAN (Ours) (CNN, FRCNN-ResNet)	**37.0**	**66.6**	**78.0**	**47.8**	**81.2**	**90.4**

Table 4.6: Performance comparison on MS-COCO's 1K and 5K testing sets. For each baseline, the best single model with highest R@1 in text-to-image retrieval task is reported and compared.

original papers are not compared for fairness.

The full model includes object-oriented encoders and inter- and intra-modal attention networks. As analyzed with the ablation studies, I choose the full model trained with the loss function emphasizing paired intra-modal hard negative examples (Eq. 4.9 and Eq. 4.11). Table 4.5 shows the quantitative results on Flickr30K.

Since most models employ RNN as the textual encoder, the trend shows that models with more advanced visual features generally perform better. Interestingly, there are clear performance gaps

when swapping the visual backbone from AlexNet (Krizhevsky et al., 2012) to VGG (Simonyan and Zisserman, 2014) or ResNet (He et al., 2016b). Sharing a similar object-oriented visual encoder (Faster RCNN) as SCAN (Lee et al., 2018), my model achieves superior performance to other models. Note that even with only ResNet backbone, my model still achieves comparable or superior performance to baselines with ResNet backbones, as shown in Table 4.4. With the object-oriented textual and visual encoders associated with inter- and intra-modal attention networks, my model achieves new state-of-the-art results on Flickr30K. OAN outperforms the previous best model (SCAN (Lee et al., 2018)) by 7.5 (16.3%), 5.7 (7.6%), 4.1 (4.9%) and 6.8 (11%), 5.5 (6.2%), 2.3 (2.4%) in R@1, R@5 and R@10 for the text-to-image and the image-to-text retrieval tasks, respectively.

MS-COCO is roughly four times larger than Flickr30K. Following the protocol in (Karpathy and Fei-Fei, 2015), I report the evaluation results on the 1K testing images (5-fold) and 5K testing images. The quantitative results of the two splits on MS-COCO is shown in Table 4.6. On both testing splits, the proposed model delivers the best cross-modal retrieval performance on most of the metrics.

4.4 Summary

In this chapter, I introduced an analysis based on adversarial perturbation with random shuffling and leave-one-POS-out to investigate and quantify what cross-modal retrieval models learn. The results show that the regional alignments between textual phrases and visual objects play an essential role in image-text matching while linguistic information such as syntactic structure and some lexical types was not well-captured even with RNN encoders.

Based on this insight, I proposed a novel model that employs object-oriented encoders along with inter- and intra-modal attention networks to pool and align visual objects and text phrases both regionally and globally. A new inter- and intra-modal structure-preserving loss function has also been introduced to enhance the intra-modal discrepancies further. In the empirical evaluation, I validated the effectiveness of the proposed model on the Flickr30K dataset and the MS-COCO dataset with state-of-the-art results at that time.

Chapter 5

Multilingual Multimodal Representation Learning

5.1 Overview

In the previous chapter, for English-image embeddings, I have presented my adversarial probing analysis which sheds insights for learning improved object-oriented embeddings. As most multimodal datasets are English-centered, it is unclear whether the success of English-vision models can be transferred to other languages. In this chapter, besides English, I follow the emerging research trend that attempts to generalize learning visual-semantic embeddings (VSE) in the multilingual scenario.

Some prior work has been proposed along the line. Rajendran et al. (2016) learn multi-view representations when parallel data is available only between one pivot view and the rest of the views. PIVOT (Gella et al., 2017) extends the work from Calixto et al. (2017) to use images as the pivot view for learning multilingual multimodal representations. Kádár et al. (2018) further confirms the benefits of multilingual training.

In this chapter, I explore efficient and effective approaches that learn improved multilingual multimodal representations (*i.e.*, multilingual-VSE). In section §5.2, for multilingual text-image retrieval, I propose to improve multilingual text-image representations with diverse multi-head attention, which captures the different types of multilingual multimodal contents. I then evaluate the learned multilingual-VSE on English and multilingual text-to-image retrieval. My work improves prior work by (a) introducing a model with multi-head attention to distill only the important information for end tasks, and (b) utilizing object/entity-level representations compared

to whole-image and sentence-level embeddings as in prior work, and (c) incorporating a novel diversity objective to improve the learned representations. In short, I verify the following benefits of learning multilingual multimodal representations:

1. Multilingual multimodal representations are more robust and generalized as the model learns visual-linguistic content from various languages.
2. Multilingual multimodal representations reason over word meanings in different languages with visual evidence. Learning these representations provides a new degree of freedom for model grounding and promotes interpretability.

The detailed content in this chapter can be found in:

1. "Multi-Head Attention with Diversity for Learning Grounded Multilingual Multimodal Representations," Po-Yao Huang, Xiaojun Chang, Alexander Hauptmann, *EMNLP 2019*

5.2 Diversified Multilingual VSE for Retrieval

5.2.1 Motivation

To promote and understand the multilingual version of image search, I leverage visual object detection and propose a model with diverse multi-head attention to learn grounded multilingual multimodal representations. Specifically, our model attends to different types of textual semantics in two languages and visual objects for fine-grained alignments between sentences and images. I introduce a new objective function that explicitly encourages attention diversity to learn an improved visual-semantic embedding space. I evaluate our model in the German-Image and English-Image matching tasks on the Multi30K dataset, and in the Semantic Textual Similarity task with the English descriptions of visual content. Results show that our model yields a significant performance gain over other methods in all of the three tasks.

5.2.2 Prior Work

An emerging trend generalizes learning VSE in the multilingual scenario. Rajendran et al. (2016) learn multi view representations when pivoted data is available. PIVOT (Gella et al., 2017), one of the earliest works studying multilingual multimodal representation learning, explores images as the pivot view. Kádár et al. (2018) further confirms the benefits of multilingual training. These prior works all focus on whole sentence-level and image-level alignments via average-pooling of

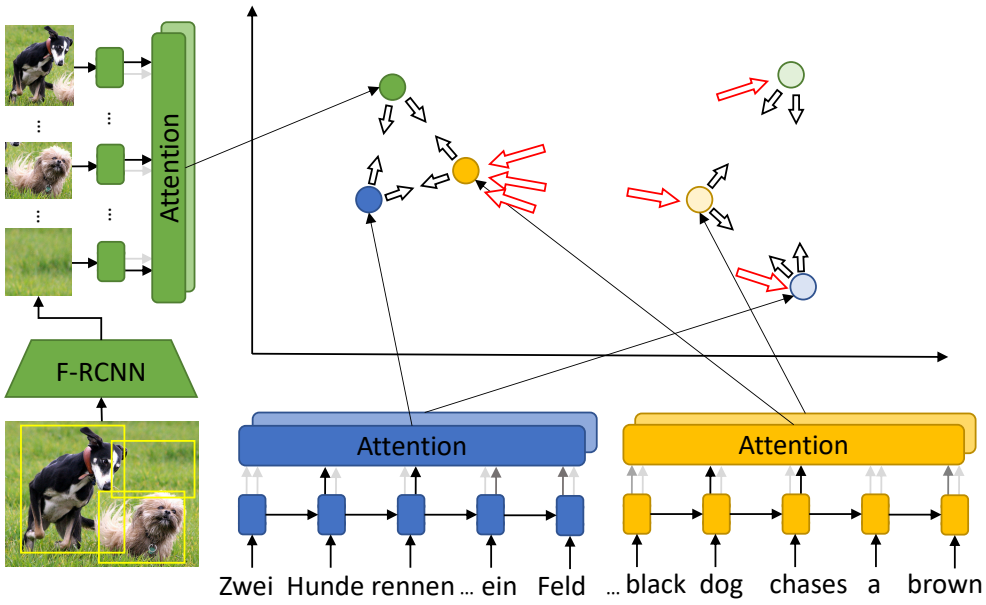


Figure 5.1: Multi-head attention with diversity for learning grounded multilingual multimodal representations. (A two-headed example with a part of diversity loss l_θ^D colored in red.)

feature map (visual part) or word embeddings (textual part), or simply use the last hidden state of CNN (visual part) or RNN (textual part).

My work is motivated by Gella et al. (2017) but has important differences. First, to disentangle the alignments in the joint embedding space, I employ visual object detection and multi-head attention to selectively align salient visual objects with textual phrases, resulting in visually-grounded multilingual multimodal representations. Second, as multi-head attention Vaswani et al. (2017) is appealing for its efficiency and ability to jointly attend to information form different perspectives, we propose to further encourage the diversity among attention heads to learn an improved visual-semantic embedding space.

5.2.3 Diversified Multi-Head Attention

Figure 5.1 illustrates the overview of the proposed model. Given a set of images as the pivoting points with the associate English and German[1] descriptions or captions, the proposed VSE model aims to learn a multilingual multimodal embedding space in which the encoded representations $(\mathbf{v}, \mathbf{e}, \mathbf{g})$ of a paired instance are closely aligned to each other than non-paired ones.

Encoders: For a sampled pair, I first encode the tokens in the English sentence $\mathbf{x}^e = \{x_1^e, \ldots, x_N^e\}$

[1]For clarity in notation, I discuss only two languages. The proposed model can be intuitively generalized to more languages by summing additional terms in Eq. 5.4 and Eq. 5.6-5.7.

and the tokens in the German sentence $\mathbf{x}^g = \{x_1^g, \ldots, x_N^g\}$ through the word embedding matrices followed by two bi-directional LSTMs. The outputs of the textual encoders are $\mathbf{e} = \{e_1, \ldots, e_N\}, e_n \in \mathbb{R}^H$ for English and $\mathbf{g} = \{g_1, \ldots, g_N\}, g_n \in \mathbb{R}^H$ for German, where N is the max sentence length and H is the dimension of the shared embedding space. For the image, I leverage a Faster-RCNN (Ren et al., 2015) network with a ResNet (He et al., 2016b) backbone to detect and encode salient visual objects in the image. With a trainable one-layered perceptron to transform visual features into the shared embedding space, I encode the image as $\mathbf{v} = \{v_1, \ldots, v_M\}, v_m \in \mathbb{R}^H$, where M is the maximum amount of visual objects in an image.

Multi-head attention with diversity: I employ K-head attention networks to attend to the visual objects in an image as well as the textual semantics in a sentence then generate fixed-length image/sentence representations for alignment. Specifically, the k-th attended German sentence representation g^k is computed by:

$$a_i^k = \text{softmax}(\tanh(\mathbf{W}_{\mathbf{c}_g}^k \mathbf{c}_g^k)^\top \tanh(\mathbf{W}_g^k \mathbf{g}_i)) \tag{5.1}$$

$$\mathbf{g}^k = \sum_{i=1}^{N} a_i^k \mathbf{g}_i, \tag{5.2}$$

where a_i^k is the attention weight, $\mathbf{W}_g^k \in \mathbb{R}^{H \times H_{attn}}$, $\mathbf{W}_{c_g}^k \in \mathbb{R}^{H_c \times H_{attn}}$ is the learnable transformation matrix for German. $\mathbf{c}_g^k \in \mathbb{R}^{H_c}$ is the learnable H_c-dimensional contextual vector for distilling important semantics from German sentences. The final German sentence representation is the concatenation of K-head attention outputs $\mathbf{g} = [\mathbf{g}^0 \| \ldots \| \mathbf{g}^K]$. Similar for encoding the English sentence $\mathbf{e} = [\mathbf{e}^0 \| \ldots \| \mathbf{e}^K]$ and the image $\mathbf{v} = [\mathbf{v}^0 \| \ldots \| \mathbf{v}^K]$.

With $\{V, E, G\}$ where $\mathbf{v} \in V, \mathbf{e} \in E, \mathbf{g} \in G$ as the set of attended fixed-length image and sentence representations in a sampled batch, I use the widely-used hinge-based triplet ranking loss with hard negative mining (Faghri et al., 2018) to align instances in the visual-semantic embedding space. Taking Image-English instances $\{V, E\}$ as an example, I leverage the triplet correlation loss defined as:

$$\begin{aligned} l_\theta(V, E) = &\sum_p \left[\alpha - s(\mathbf{v}_p, \mathbf{e}_p) + s(\mathbf{v}_p, \hat{\mathbf{e}}_p)\right]_+ \\ &+ \sum_q \left[\alpha - s(\mathbf{v}_q, \mathbf{e}_q) + s(\hat{\mathbf{v}}_q, \mathbf{e}_q)\right]_+, \end{aligned} \tag{5.3}$$

where α is the correlation margin between positive and negative pairs, $[.]_+ = \max(0, .)$ is the hinge function, and $s(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a}^T \mathbf{b}}{\|\mathbf{a}\| \|\mathbf{b}\|}$ is the cosine similarity. p and q are the indexes of the images and sentences in the batch. $\hat{\mathbf{e}}_p = \text{argmax}_q s(\mathbf{v}_p, \mathbf{e}_{q \neq p})$ and $\hat{\mathbf{v}}_q = \text{argmax}_p s(\mathbf{v}_{p \neq q}, \mathbf{e}_q)$ are the hard negatives. When the triplet loss decreases, the paired images and German sentences are drawn

closer down to a margin α than the hardest non-paired ones. Our model aligns $\{V, E\}$, $\{V, G\}$ and $\{E, G\}$ in the joint embedding space for learning multilingual multimodal representations with the sampled $\{V, E, G\}$ batch. I formulate the overall triplet loss as:

$$l_\theta(V, E, G) = l_\theta(V, G) + l_\theta(V, E) + \gamma l_\theta(G, E). \tag{5.4}$$

Note that the hyper-parameter γ controls the contribution of $l_\theta(G, E)$ since ($\mathbf{e}$, $\mathbf{g}$) may not be a translation pair even though ($\mathbf{e}$, $\mathbf{v}$) and ($\mathbf{g}$, $\mathbf{v}$) are image-caption pairs.

One of the desired properties of multi-head attention is its ability to jointly attend to and encode different information in the embedding space. However, there is no mechanism to support that these attention heads indeed capture diverse information. To encourage the diversity among K attention heads for instances within and across modalities, I propose a new simple, yet effective margin-based diversity loss. As an example, the multi-head attention diversity loss between the sampled images and the English sentences (*i.e.* diversity across-modalities) is defined as:

$$l_\theta^D(V, E) = \sum_p \sum_k \sum_r \left[\alpha_D - s(\mathbf{v}_p^k, \mathbf{e}_p^{k \neq r})\right]_+ \tag{5.5}$$

As illustrated with the red arrows for the update in Figure 5.1, the merit behind this diversity objective is to increase the distance (up to a diversity margin α_D) between attended embeddings from different attention heads for an instance itself or its cross-modal parallel instances. As a result, the diversity objective explicitly encourages multi-head attention to concentrate on different aspects of information sparsely located in the joint embedding space to promote fine-grained alignments between multilingual textual semantics and visual objects. With the fact that the shared embedding space is multilingual and multimodal, for improving both intra-modal/lingual and inter-modal/lingual diversity, I model the overall diversity loss as:

$$\begin{aligned} l_\theta^D(V, E, G) = l_\theta^D(V, V) + l_\theta^D(G, G) + l_\theta^D(E, E) \\ + l_\theta^D(V, E) + l_\theta^D(V, G) + l_\theta^D(G, E), \end{aligned} \tag{5.6}$$

where the first three terms are intra-modal/lingual and the rest are cross-modal/lingual. With Eq. 5.4 and Eq. 5.6, I formalize the final model loss as:

$$l_\theta^{All}(V, E, G) = l_\theta(V, E, G) + \beta l_\theta^D(V, E, G), \tag{5.7}$$

where β is the weighting parameter which balances the diversity loss and the triple ranking loss. I train the model by minimizing $l_\theta^{All}(V, E, G)$.

5.2.4 Empirical Evaluation

Experiment Setup

I use the model in Anderson et al. (2018), which is a Faster-RCNN (Ren et al., 2015) network pre-trained on the MS-COCO (Lin et al., 2014) dataset and fine-tuned on the Visual Genome (Krishna et al., 2017b) dataset to detect salient visual objects and extract their corresponding features. 1,600 types of objects are detectable. I then pack and represent each image as a 36×2048 feature matrix where 36 is the maximum amount of salient visual objects in an image, and 2048 is the dimension of the flattened last pooling layer in the ResNet (He et al., 2016b) backbone of Faster-RCNN.

For the text processing, I lower-case, tokenize, and then truncate the maximum sentence length to 100. I use 300-dim word embedding matrices initialized either randomly or with pre-trained multilingual embeddings. (I use the multilingual version of FastText (Mikolov et al., 2018)). I also experiment incorporating the last layer of contextualized multilingual BERT embeddings (Devlin et al., 2019a) to replace the word embedding matrices as the textual input features for the bi-directional LSTMs.

For training, I sample batches of size 128 and train 20 epochs on the training set of Multi30K. I use the Adam (Kingma and Ba, 2014) optimizer with 2×10^{-4} learning rate then 2×10^{-5} after 15-*th* epoch. Models with the best summation of validation R@1,5,10 are selected to generate the image and sentence embeddings for testing. Weight decay is set to 10^{-6}, and gradients larger than 2.0 are clipped. I use 3-head attention ($K = 3$) and the embedding dimension $H = 512$. All the context vectors share the same dimension in the attention modules. Other hyper-parameters are set as follows: $\alpha = 0.2, \alpha_D = 0.1, \beta = 1.0$ and $\gamma = 0.6$.

Multilingual Sentence-Image Matching

I evaluate the proposed model in the multilingual sentence-image matching (retrieval) tasks on Multi30K: (i) Searching images with text queries (Sentence-to-Image). (ii) Ranking descriptions with image queries (Image-to-Sentence). English and German are considered. Two types of annotations are available in Multi30K: (i) One parallel English-German translation for each image and (ii) five independently collected English and five German descriptions/captions for each image. I use the later. Note that the German and English descriptions are *not* translations of each other and may describe an image differently.

Table 5.1 presents the results on the Multi30K testing set. The VSE baselines in the first five rows are trained with English and German descriptions independently. In contrast, PIVOT (Gella

Method	German to Image			Image to German			English to Image			Image to English		
	R@1	R@5	R10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R10
VSE$^{\dagger *}$ (Kiros et al., 2014)	20.3	47.2	60.1	29.3	58.1	71.8	23.3	53.6	65.8	31.6	60.4	72.7
OE$^{\dagger *}$ (Vendrov et al., 2015)	21.0	48.5	60.4	26.8	57.5	70.9	25.8	56.5	67.8	34.8	63.7	74.8
DAN* (Nam et al., 2017)	31.0	60.9	71.0	46.5	77.5	83.0	39.4	69.2	69.1	55.0	81.8	89.0
VSE++* (Faghri et al., 2018)	31.3	62.2	70.9	47.5	78.5	84.5	39.6	69.1	79.8	53.1	82.1	87.5
SCAN* (Lee et al., 2018)	35.7	64.9	74.6	52.3	81.8	88.5	45.8	74.4	83.0	61.8	87.5	93.7
Pivot† (Gella et al., 2017)	22.5	49.3	61.7	28.2	61.9	73.4	26.2	56.4	68.4	33.8	62.8	75.2
Ours† (Rand+VGG)	25.8	54.9	65.1	34.1	65.5	76.5	30.1	62.5	71.6	36.4	68.0	80.9
Ours (No Diversity)	36.3	65.3	74.7	53.1	82.3	88.8	46.2	74.7	82.9	63.3	87.0	93.3
Ours (Rand)	39.2	67.5	76.7	55.0	84.7	91.2	48.7	77.2	85.0	66.4	88.3	93.4
Ours (w/ FastText)	40.3	70.1	**79.0**	**60.4**	**85.4**	**92.0**	**50.1**	78.1	85.7	**68.0**	88.8	94.0
Ours (w/ BERT)	**40.7**	**70.5**	78.8	56.5	84.6	91.3	48.9	**78.3**	**85.8**	66.5	**89.1**	**94.1**

Table 5.1: Comparison of multilingual sentence-image retrieval/matching (German-Image) and (English-Image) results on Multi30K. (Visual encoders:VGG† otherwise ResNet or Faster-RCNN(ResNet).) (Monolingual models*.)

Figure 5.2: Qaulitative text-to-image matching results on Multi30K. Correct (colored in green) if ranked first.

Figure 5.3: t-SNE visualization and grounding of the learned multilingual multimodal embeddings on Multi30K. Note the sentences are *not* translation pairs.

et al., 2017) and the proposed model are capable of handling multilingual input queries with a single model. For a fair comparison with PIVOT, I also report the result of swapping Faster-RCNN with VGG as the visual feature encoder in our model.

As can be seen, the proposed models successfully obtain state-of-the-art results, outperforming other baselines by a significant margin. German-Image matching benefits more from the joint training with English-Image pairs. The models with pre-trained multilingual embeddings and contextualized embeddings achieve better performance in comparison to randomly initialized word embeddings, especially for German. One explanation is that the multi-task training alleviates the degradation from German singletons with English and the pre-trained embeddings. While the model with BERT performs better in English, FastText is preferred for German-Image matching.

Qualitative Results and Grounding

In Figure 5.2, I sample some qualitative multilingual text-to-image matching results. In most cases, our model successfully retrieves the one and only one correct image. Figure 5.3 depicts the t-SNE visualization of the learned visually grounded multilingual embeddings of the (v, e, g) pairs pivoted on v in the Multi30K testing set. As evidenced, although the English and German sentences describe different aspects of the image, our model correctly aligns the shared semantics (*e.g.* ("*man*", "*mann*"), ("*hat*", "*wollmütze*")) in the embedding space. Notably, the embeddings are visually-grounded as our model associate the multilingual phrases with exact visual objects (*e.g. glasses* and *ears*). I consider learning a grounded multilingual multimodal dictionary as the promising next step.

As limitations, I notice that actions and small objects are harder to align. Additionally, the alignments tends to be noun-phrase/object-based whereas spatial relationships (*e.g.* "*on*", "*over*") and quantifiers remain not well-aligned. Resolving these limitations will be our future work.

5.3 Summary

Multilingual multimodal representation learning is challenging yet rewarding in many multilingual vision-language tasks. In this chapter, I have presented methods of learning multilingual image-text representations for multilingual cross-modal retrieval.

I presented a novel model that facilitates multi-head attention with diversity to align different types of textual semantics and visual objects for learning grounded multilingual multimodal representations. The proposed model obtains state-of-the-art results at that time in the multilingual sentence-image matching task and the semantic textual similarity task on two benchmark datasets.

Chapter 6

Multilingual Multimodal Pre-training at Scale

6.1 Overview

In Chapter §4 and Chapter §5, my work in this thesis has pushed multiple research fronts in monolingual multimodal and multilingual multimodal representation by leveraging image object detection, multi-head attention with diversity, and proper neural architectures. Some progress has been made on small (30∼120K) and well-annotated image-based datasets such as Flick30K, Multi30K, and MS-COCO. However, there are still many remaining challenges and opportunities, in particular, pre-training with large-scale noisy user-generated videos, which has attracted much research attention in recent years. Following this trend, in this chapter, I focus on two important topics: (1) learning from million-scale noisy user-generated videos and (2) exploiting multilingual captions for multilingual text-video representation learning.

6.1.1 Challenges in Video-Text Representation Learning

Nowadays, it is feasible to access hundreds of millions of videos and their corresponding captions in different languages uploaded by the users on popular social media platforms such as YouTube, Twitter, and TikTok. With the unprecedented amount of multi-million-scale videos, scalable and efficient learning methods that would cover these multimodal data would be desirable. Although videos are a popular type of user-generated multimedia content, there are much fewer research works focus on encoding video and text contents because of the following challenges: First, videos are harder to process and encode compared to images. For example, a caption may correspond

to a 10-second video in MSR-VTT that contains 300 frames. In comparison, there is only 1 image (1 frame) in Flick30K or MS-COCO. Second, Video captions are ambiguous as there is an additional temporal domain to be annotated. As a result, the annotations in most video-text datasets are usually noisier and more ambiguous as the annotators tend to focus on different temporal segments and annotate differently. Lastly, The automated subtitles via automated speech recognition (ASR) in popular video-text datasets are very noisy. For example, only the subtitles generated by Google ASR are provided for the instructional videos in the HowTo100M dataset (see Chapter §3 for details). Consequently, besides the misalignment of the speech (instruction) and the action (demonstration) by the performers in these instructional videos, there is inherent noise from ASR models.

6.1.2 Towards Multilingual Vision-Language Models

Besides the challenges for modeling videos, the characteristics of video-text data also provide a unique opportunity for broadening the impact of various vision-language models. Much recent progress in vision-language applications is driven by the availability of large-scale multimodal data. However, existing multimodal datasets are mostly English-based, and thus limit their applicability to generalize to non-English languages. Consequently, developing methods that could generalize English-based vision-language models to non-English languages would be a crucial step towards generalizing these models that would have a broader impact on non-English speakers.

For video-text data in the popular social media platforms mentioned above, in addition to the English captions, there are many off-the-shelf user-generated captions or machine-translated in various languages available. Therefore, these multilingual text-video data provide a unique opportunity for learning multilingual multimodal representations. As in multilingual pre-training for cross-lingual transfer in NLP applications, I envision that multilingual multimodal pre-training would facilitate cross-lingual transfer for vision-language models.

6.1.3 Chapter Organization

In this chapter, I firstly investigate and improve self-supervised learning of English-video representations from millions of noisy instructional videos and their transcriptions. Then I develop a task-agnostic multilingual text-video pre-training strategy to leverage multilingual captions for cross-lingual transfer. This chapter is organized as the following:

1. Large-scale task-agnostic pre-training for learning English-video representations. (§6.2)
2. Cross-lingual transfer of vision-language models via multilingual text-video pre-training. (§6.3)

In Section §6.2, I propose a novel method to improve self-supervised learning of video-text representations at scale. Conventionally, the dominant learning paradigm, noise contrastive learning, increases the similarity of the representations of pairs of samples that are known to be related, such as text and video from the same sample, and their artificially-transformed versions. I posit that this last behavior is too strict, enforcing dissimilar representations even for samples that are semantically related – for example, visually similar videos or ones that share the same depicted action. To this end, I propose a natural way to push semantically-related samples together without any supervision: each sample's representation must be reconstructed as a weighted combination of other support samples. Empirically, the proposed method achieves strong state-of-the-art retrieval results across common video-text benchmark tasks.

Base on the progress above, in Section §6.3, I then address the problem of *zero-shot cross-lingual transfer* of vision-language models with the learned multilingual multimodal representation. To achieve this goal, I extend the instructional video collection of the HowTo100M (Miech et al., 2019) dataset to collect transcriptions in 9 languages and construct the Multi-HowTo100M dataset, the largest multilingual text-video dataset (see Appendix for details). I then leverage tailored cross-modal cross-lingual noise contrastive objectives to pre-train multilingual multimodal embeddings at scale. After English-only task-specific fine-tuning, I evaluate the cross-lingual transferability of the fine-tuned vision-language model on the target language without using additional annotation. My key findings suggest that compared to the text-only cross-lingual transfers of NLP models (Hu et al., 2020), there is a much limited zero-shot transferability for vision-language models with text-pretrained multilingual Transformers. The results show that the proposed multilingual multimodal pre-training is the key step to generalize vision-language models across various languages.

The content in this chapter appears in:

1. "Support-set bottlenecks for video-text representation learning," Mandela Patrick*, Po-Yao Huang*, Yuki Asano*, Florian Metze, Alexander Hauptmann, João Henriques, Andrea Vedaldi, *ICLR 2021*
2. "Multilingual Multimodal Pre-training for Zero-Shot Cross-Lingual Transfer of Vision-Language Models," Po-Yao Huang*, Mandela Patrick*, Junjie Hu, Graham Neubig, Florian Metze, Alexander Hauptmann, *NAACL 2021*.

6.2 Bottlenecks in Video-Text Representation Learning

6.2.1 Motivation

With the exploding amount of user-uploaded videos available, modeling videos and learning joint video-text representations for end tasks such as text-to-video search, video captioning, and video question answering have attracted much research attention. For learning video-text representation, one method which gains increasing popularity is noise contrastive learning (Gutmann and Hyvärinen, 2010) that learns data representations both for supervised (Khosla et al., 2020) and unsupervised regimes (Chen et al., 2020b). The idea is to learn a representation that discriminates any two data samples while being invariant to certain data transformations. For example, one might learn a representation that identifies a specific image up to arbitrary rotations (Misra and van der Maaten, 2020). In a multi-modal setting, the transformations can separate different modalities, for example, by extracting the text and visual signals from a video.

The noise contrastive approach is motivated by the fact that the transformations that are applied to the data samples leave their 'meaning' unchanged. For example, rotating an image does not change the fact that it contains a cat or not (Gidaris et al., 2018). However, in most cases, I expect to find many data samples that share the same content without being necessarily related by simple transformations (*i.e.*, think of any two images of cats). Existing noise contrastive formulations are unaware of these relationships and still try to assign different representations to these samples (Wu et al., 2018), despite the fact that they are semantically equivalent. If the representation is learned for a downstream task such as semantic video retrieval, this might degrade performance.

This suggest that there might be other learning signals that could complement and improve pure contrastive formulations. In this paper, I explore this idea in the case of learning from two modalities: videos and text, in the form of video transcripts or captions. Given a state-of-the-art contrastive formulation that learns from these two modalities, I investigate complementary pretext objectives to improve it. First, I consider the *(instance) captioning* task, namely mapping a video to the corresponding text, casting this as a conditional stochastic text generation problem. I show that this brings only a modest benefit.

I observe that the captioning task is highly sample-specific, as the goal is to produce a caption which describes a specific video and not any other video, and thus it suffers from the same disadvantages (discouraging concept sharing among samples) as contrastive learning. Thus, I propose to address this issue by switching to a different text generation task. The idea is to modify the text generator to take as input a learnable mixture of a support-set of videos, which I call

cross-instance captioning. The mixture weights are generated by comparing the learned video representations to captions' representations in an online way over the batch. The limited set of support samples acts as a bottleneck that encourages extraction of shared semantics. In this manner, the embeddings can associate videos that share similar captions even if the contrastive loss tries to push them apart.

I show that, when the captioning task is added in this manner, it brings a sensible improvement to already very strong video representation learning results, further improving my own state-of-the-art baseline by a significant margin.

6.2.2 Prior Work

Large-scale Video-Text Representation Learning Large-scale training data has enabled the more effective pretraining of image (Yalniz et al., 2019; Sun et al., 2017), video (Ghadiyaram et al., 2019; Thomee et al., 2016) and textual representations (Raffel et al., 2019). The release of the HowTo100M dataset (Miech et al., 2019), a large-scale instructional video dataset, has spurred significant interest in leveraging large-scale pre-training to improve video-text representations for tasks such as video question-answering (Lei et al., 2018), text-video retrieval (Liu et al., 2019) and video captioning (Zhou et al., 2018b) on smaller datasets such as YouCookII (Zhou et al., 2018a), MSVD (Venugopalan et al., 2015), MSR-VTT (Xu et al., 2016), LSMDC (Rohrbach et al., 2017), DiDeMo (Hendricks et al., 2018) and ActivityNet (Krishna et al., 2017a). Although semantically rich and diverse, instructional videos from the web are super noisy and therefore a few approaches have been proposed to combat this. A few works (Sun et al., 2019b,a; Zhu and Yang, 2020; Luo et al., 2020) extend the BERT model to accept both visual and textual tokens to learn high-level semantic video-text representations. Other works have leveraged the contrastive loss (Miech et al., 2020) and show that using the raw audio (Rouditchenko et al., 2020; Alayrac et al., 2020) and other modalities (Gabeur et al., 2020) can be used to better align and improve video-text representations. While all these approaches rely on a contrastive objective, VidTranslate (Korbar et al., 2020) shows that a generative objective can also be used to learn joint video-text representations. In contrast to (Korbar et al., 2020), I show that combining contrastive and generative objectives to pre-train video-text representations on large-scale data such as HowTo100M is very effective. The generative objective serves as regularizer to mitigate the strictness of the instance discrimination task of the constrastive objective, showing benefits similar to approaches such as clustering (Li et al., 2020; Caron et al., 2020) and feature mixing (Kalantidis et al., 2020) which have been applied in the image domain.

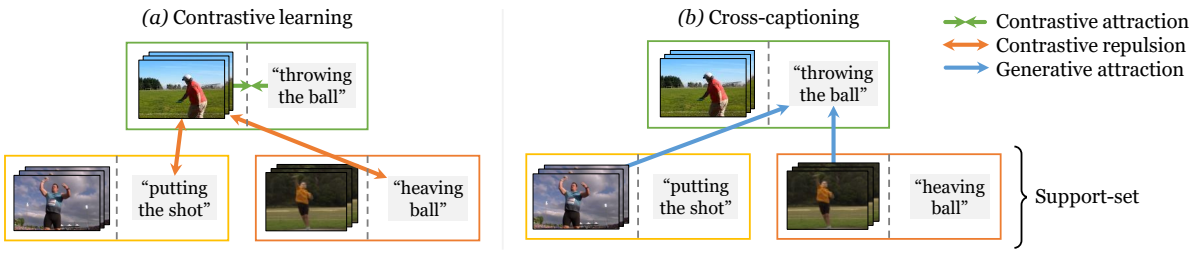


Figure 6.1: **Cross-modal discrimination and cross-captioning.** My model learns from two complementary losses: (a) Cross-modal contrastive learning learns strong joint video-text embeddings, but every other sample is considered a negative, pushing away even semantically related captions (orange arrows). (b) I introduce a generative task of cross-captioning, which alleviates this by learning to reconstruct a sample's text representation as a weighted combination of a support-set, composed of video representations from other samples.

6.2.3 Contrastive Learning with Generative Objectives

I consider the problem of learning multimodal representations from a corpus $\mathcal{C}$ of video-text pairs (v, t), where v is a video and t is its corresponding text (caption or transcription). My goal is to learn a pair of representation maps $c_v = \Psi(v)$ and $c_t = \Phi(t)$, with outputs in a d-dimensional embedding space $c_v, c_t \in \mathbb{R}^d$, where semantically similar instances are close to each other.

Objective for Learning Multimodal Representations

I consider two learning objectives, also illustrated in Figure 6.1. The first is the contrastive objective, pushing embeddings c_t and c_v to be close if text t and video v come from the same sample and pushing them apart otherwise. This assumes that every sample is its own class and does not benefit from modelling similiarities *across* instances. The second objective is generative captioning. In its most basic variant, it maximizes the probability of generating the text t given the corresponding video v. However, I suggest that variants that explicitly promote concept sharing between instances will result in better downstream performance, in tasks such as video retrieval. These variants, illustrated in Figure 6.2, have in common that the caption t is reconstructed from a learned weighted combination over *other* videos $\hat{v}$. This is a form of attention (Bahdanau et al., 2015) which encourages the network to learn about which videos share similar semantics, compensating for the contrastive loss and grouping them implicitly.

In the following, I denote with $\mathcal{B} \subset \mathcal{C}$ a *batch* of multi-modal samples, i.e. a finite collection of video-text pairs $(t, v) \in \mathcal{C}$. For simplicity, I denote the batch as $\mathcal{B} = \{(t^i, v^i)\}_{i=1}^{B}\}$.

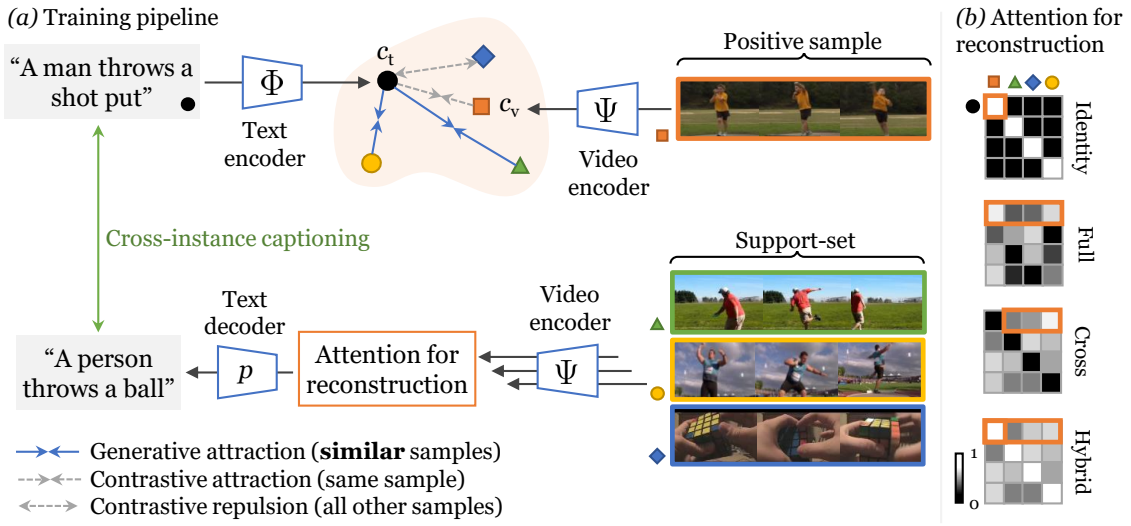


Figure 6.2: **(a)** My cross-modal framework with the discriminative (contrastive) objective and the generative objective. The model learns to associate video-text pairs in a common embedding space with text and video encoders (top). Meanwhile, the text must also be reconstructed as a weighted combination of video embeddings from a support-set (bottom), selected via attention, which enforces representation sharing between different samples. **(b)** Weights matrices (attention maps) used in each cross-captioning objective (see section 6.2.3).

Contrastive objective To define the contrastive objective, let $s(a, b) = \frac{a^\top b}{\|a\|\|b\|}$ be the similarity measure between vectors a and b. Following (Faghri et al., 2018), I adopt the hinge-based triplet ranking loss with hard negative mining:

$$\mathcal{L}^{\text{contrast}} = \frac{1}{B}\sum_{i=1}^{B}\left[\max_j \left[\alpha - s(c_t^i, c_v^i) + s(c_t^i, c_v^j)\right]_+ + \max_j \left[\alpha - s(c_t^i, c_v^i) + s(c_t^j, c_v^i)\right]_+\right], \tag{6.1}$$

where α is the correlation margin between positive and negative pairs and $[\cdot]_+ = \max\{0, \cdot\}$ is the hinge function. In the experiments, I set $\alpha = 0.2$.

Cross-captioning objectives In the conventional captioning, the decoder seeks to optimize the negative log-likelihood of a text sequence t given its corresponding video v:

$$\mathcal{L}^{\text{caption}} = -\frac{1}{B}\sum_{i=1}^{B}\log p(t^i|e_v^i). \tag{6.2}$$

Here, the log-likelihood is obtained via auto-regressive decoding (Vaswani et al., 2017) from an intermediate video embedding $e_v^i = \Phi'(v^i)$. For the cross-captioning objective, I modify this loss

to condition the generation process on a weighted average of the embeddings of the *other* videos in the batch, which I call the *support-set*. The weights themselves, which can be interpreted as a batch-wise attention, are obtained as a softmax distribution with temperature T over batch indices based on the video embeddings, as follows:

$$\mathcal{L}^{\text{cross-captioning}} = -\frac{1}{B}\sum_{i=1}^{B} \log p(t^i|\bar{e}_v^i), \quad \bar{e}_v^i = \sum_{j\in\mathcal{S}_i} \frac{\exp\langle c_t^i, c_v^j\rangle/T}{\sum_{k\in\mathcal{S}_i}\exp\langle c_t^i, c_v^k\rangle/T}\cdot e_v^j. \tag{6.3}$$

By default, the summation in the softmax is conducted over a support set $\mathcal{S}_i$ containing all indices except i. In the experiments, I consider the following attention types for reconstruction. **Identity captioning** ($\mathcal{S}_i = \{i\}$) generates the caption from the corresponding video and reduces to the standard captioning objective, eq. (6.2). **Full support** ($\mathcal{S}_i = \{1, \ldots, B\}$) considers all videos as possible candidates for captioning. **Hybrid captioning** sets the weights in eq. (6.3) as the average of the weights for identity captioning and full support. **Cross-captioning** ($\mathcal{S}_i = \{j \neq i\}$) considers all *but* the video that one wishes to caption. This variant forces the network to extract all information required for captioning from other videos in the batch. Figure 6.2 compares graphically these attention mechanisms.

Considering both discriminative and generative objectives for learning multimodal representations, the full objective is $\mathcal{L} = \mathcal{L}^{\text{contrast}} + \lambda\mathcal{L}^{\text{cross-captioning}}$, where λ balances two objectives. I set $\lambda = 10$ to ensure similar magnitudes for both losses in the experiments. In the training phase, I use Adam (Kingma and Ba, 2014) to minimize the loss. At inference time, I directly use $\Phi(t)$ and $\Psi(v)$ to encode video and text representations for retrieval.

Model Architecture I now discuss the details of the encoders and decoder components in the architecture, illustrated in fig. 6.2. For the *text decoder* $p(t|e_v)$ in eqs. (6.2) and (6.3), I use a pre-trained T-5 decoder (Raffel et al., 2019).

For the *video representation* $c_v = \Psi(v) = \Psi''(\Psi'(v))$, I use a video encoder $e_v = \Psi'(v)$ followed by a multi-layer transformer pooling head $c_v = \Psi''(e_v)$. The encoder $\Psi'(v)$ concatenates the output of pretrained ResNet-152 (He et al., 2016b) and R(2+1)D-34 (Tran et al., 2018) networks applied to individual video frames, resulting in a code $e_v = [e_{v1} \cdots e_{vM}]$ where M is the maximum duration of a video clip. For the pooling head $c_v = \Psi''(e_v)$, I consider a transformer architecture to attend to important context and summarize it into a fixed-length representation c_v. For this, I follow MMT (Gabeur et al., 2020), but with two important differences. First, while MMT uses 7 expert features that results in $7\times$ the sequence length, I only use a transformer to attend to early-fused motion and appearance features as the video representation, thus significantly reducing

the sequence length and computational cost. Second, instead of stacking 6 transformer layers to encode the visual stream as in MMT, I only use a shallow two-layer transformer architecture with additional pre-encoders, further increasing model efficiency. As temporal 1D-convolutional neural networks (CNNs) (LeCun et al., 1998) were shown to effectively capture temporal dependencies in videos (He et al., 2016a), I integrate CNNs into the transformer pooling heads to better capture video temporal signals. In more detail, I compute $c_v = \Psi''(e_v)$ by chaining two transformer layers, each of the type:

$$\psi(e) = \text{BN}(\text{FFN}(e_{\text{attn}}) + e_{\text{attn}}), \quad e_{\text{attn}} = \text{BN}(\text{MHA}(f(e)) + f(e)). \tag{6.4}$$

Here f is a pre-encoder that refines the video representation; I found empirically that a 1D CNN works well for this purpose. Then, I apply multi-head self-attention (MHA) (Vaswani et al., 2017; Huang et al., 2019b) followed by a feed-forward network (FNN) with batch normalization (BN) (Ioffe and Szegedy, 2015). The architecture maps the input sequence e_v to a new 'contextualized' sequence of representation vectors; I take the first one as c_v.

The text representation decomposes in the same way as $c_t = \Phi(t) = \Phi''(\Phi'(t))$. The text encoder $e_t = \Phi'(t)$ uses a pretrained T-5 network resulting in a code $e_t = [e_{t1} \cdots e_{tN}]$, where N is the maximum length of a sentence. The pooling head $c_t = \Phi''(e_t)$ follows the same design as the video case, but f is set to a recurrent neural network (RNN) instead of a CNN. Please refer to Appendix for more details. In practice, for computational reasons, I use eq. (6.3) to finetune the parameters of all networks except the video encoder $\Psi'(v)$, which is fixed.

6.2.4 Empirical Evaluation

I validate empirically the ability of my method to learn better representations for the downstream tasks of text-to-video and video-to-text retrieval. First, I ablate various model components on the MSR-VTT dataset. Then, I show that my best model significantly outperforms state-of-the-art retrieval systems on three datasets, MSR-VTT, ActivtyNet and VATEX. Finally, I analyse qualitatively the effect of the attention mechanism used during training.

Experimental Setup

Datasets **HowTo100M** (Miech et al., 2019) is a large-scale instructional video collection of 1.2 million YouTube videos, along with automatic speech recognition transcripts. There are more than 100 million clips (ASR segments) defined in HowTo100M. I use this dataset for the pre-training experiments. **MSR-VTT** (Xu et al., 2016) contains 10,000 videos, where each video is annotated

	$R@1\uparrow$	$R@5\uparrow$	Md$R\downarrow$
None	25.9	53.0	4.0
Identity	26.4	51.9	4.0
Full	25.8	53.9	3.0
Hybrid	26.0	54.8	3.0
Cross	**27.2**	**55.2**	**3.0**

Table 6.1: **Effect of learning objectives.** Text→Video retrieval on MSR-VTT.

with 20 descriptions. I report results on the 1k-A split (9,000 training, 1,000 testing) as in (Liu et al., 2019). **VATEX** (Wang et al., 2019) is a multilingual (Chinese and English) video-text dataset with 34,911 videos. I use the official training split with 25,991 videos and report on the validation split as in HGR (Chen et al., 2020a). The **ActivityNet Caption** (Krishna et al., 2017a) dataset consists of densely annotated temporal segments of 20K YouTube videos. I use the 10K training split to train from scratch/ finetune the model and report the performance on the 5K 'val1' split. The **MSVD** (Chen and Dolan, 2011) dataset consists of $80K$ English descriptions for 1,970 videos from YouTube, with each video associated with around 40 sentences each. I use the standard split of 1,200, 100, and 670 videos for training, validation, and testing (Venugopalan et al., 2015; Xu et al., 2015; Liu et al., 2019).

Evaluation Metrics To measure the text-to-video and video-to-text retrieval performance, I report Recall at K (R@K) where $K = 1, 5, 10$. R@K is the fraction of queries that retrieve desired items in the top K candidates, where a higher score denotes a better model. I also report MdR that measures the median rank of desired items in the retrieved ranking list. These are the common metrics in the literature of information retrieval.

Ablation Studies

In Table 6.2, I first only ablate the cross-modal retrieval part of the network architecture, while the generative objectives are analysed in Table 6.1. The ablation results show that:

Video Encoder In Table 6.2a, I show the effect of the choice of visual input features. I find that for text-to-video retrieval at Recall at 1 and 5 ($R@1, R@5$), features obtained from a video R(2+1)D-34 ResNet achieve 2.9% and 7.0% higher performance compared to only image-frame based features from a ResNet-152. A further 3.5% and 2.0% can be gained by concatenating both features, yielding the strongest MdR of 3.0%.

(a) **Video Encoder**. Stronger features and combination improves performance.

Feature source	$R@1 \uparrow$	$R@5 \uparrow$	$MdR \downarrow$
R-152	20.8	46.2	6.0
R(2+1)D-34	23.7	53.2	4.0
R(2+1)D-34 +R-152	**27.2**	**55.2**	**3.0**

(b) **Feature Aggregation.** Learning temporal attention yields strong gains over pooling.

Temporal reduction	$R@1 \uparrow$	$R@5 \uparrow$	$MdR \downarrow$
Max	21.8	49.5	8.0
Mean	22.5	51.3	6.0
Multi-Head Attn	**27.2**	**55.2**	**3.0**

(c) **Text Encoder.** Stronger encoding of text improves retrieval.

Text Encoder	$R@1 \uparrow$	$R@5 \uparrow$	$MdR \downarrow$
W2V (GloVe)	22.1	49.8	6.0
T5-Small	24.5	51.2	3.0
T5-Base	**27.2**	**55.2**	**3.0**

(d) **Text Decoder.** Stronger decoding of text improves retrieval.

Text Encoder	Text Decoder	$R@1 \uparrow$	$R@5 \uparrow$	$MdR \downarrow$
T5-Base	T5-Small	26.2	54.2	3.0
T5-Base	T5-Base	**27.2**	**55.2**	**3.0**

(e) **Contrastive Loss.** Inter-modal Triplet loss yields the best performance.

Contrastive	$R@1 \uparrow$	$R@5 \uparrow$	$MdR \downarrow$
InfoNCE (inter+intra)	10.7	28.5	15.0
InfoNCE (inter)	10.8	29.0	14.5
Triplet (inter+intra)	26.8	**56.2**	3.0
Triplet (inter)	**27.2**	55.2	**3.0**

(f) **Support-set Size.** Retrieval degrades when reconstructing from too small and too large sets.

	Batch-size							**Memory bank**	
Size	8	16	32	64	128	256	512	2k	8k
R@1/5	18.5/45.6	20.7/49.9	25.2/54.6	27.2/55.2	**28.0/56.1**	26.9/55.0	25.3/53.5	26.8/54.7	26.2/52.7

Table 6.2: **Model Architecture and Training Details Ablation.** Text→Video retrieval performance on MSR-VTT. Recall@$1, 5$, and Median Recall are shown.

Feature Aggregation While the features from both video and image-based visual encoders have reduced spatial extent after a fully-connected layer, the temporal dimension can be reduced in various ways. In table 6.2b, I find that the multi-head, parameterized attention reduction yields strong gains over the mean- or max-pooling baselines of over 4% for $R@1$. This shows that learning attention over the temporal dimension of fixed feature sets can give strong gains even without fine-tuning the encoder.

Text Encoder In Table 6.2c, I find decent gains of 2.7% and 0.4% for R@1,5 for using T5-base, instead of T5-small. I do not use the T-5-Large model, as in (Korbar et al., 2020), due to the prohibitively large relative model size increase of +220%.

Text Decoder In Table 6.2d, I find that using a larger text decoder gives a 1% increase in performance when using the cross-captioning objective.

Contrastive Loss To validate the choice of a triplet loss in eq. (6.1), in table 6.2e, I compare the results of the InfoNCE contrastive loss (van den Oord et al., 2018) with a triplet loss, with both the intra and inter-intra modality variants. I find that InfoNCE (van den Oord et al., 2018) loss does not work well in my case, likely due to the difficulty in tuning this loss to have the right combination of temperature and batch-size.

Support-Set Size Lastly, in table 6.2f, I show the effect of the size of the support set used for cross-instance captioning. I find that the reconstruction loss indeed acts as a bottleneck, with both smaller and very large sizes degrading the performance.

Captioning Objective In table 6.1, I show the effect of the different variants of the learning objective eq. (6.3). First, I find that the naive addition of a reconstruction objective ("Identity") does not improve the contrastive-only baseline ("None") much. Considering reconstruction from other videos improves the performance more. In particular, the "Hybrid" variant, which combines "Identity" and "Full" (section 6.2.3) improves Recall at 1 and 5 from 25.9% and 53.0% to 26.0% and 54.8%, respectively. However, the best result by far ($27.2/55.2\%$) is obtained forcing captions to be reconstructed only from *other* videos, via the proposed cross-instance attention mechanism ("Cross"). This variant cannot use information contained in a video to generate the corresponding caption and thus entirely relies on the model to discover meaningful relationship between different videos. This newly-proposed scheme seems to have the most beneficial effect for semantic retrieval.

Comparison to State-of-the-Art

In this section, I compare the results of my method to other recent text-to-video and video-to-text retrieval approaches on various datasets. In tables 6.3 to 6.5, I show the results of my model applied to text-to-video and video-to-text retrieval on MSR-VTT, VATEX, ActivityNet and MSVD with and without pre-trainig on HowTo100M. Without pre-training, my method outperforms all others in all metrics and datasets. In particular, for the VATEX dataset, the retrieval performance at recall at 1 and 5 is 45.9% and 82.4%, exceeding recent state-of-the-art methods (Chen et al., 2020a) by a margin of 9%. For ActivityNet, my model outperforms MMT by a margin of 4%

Table 6.3: **Retrieval performance on the MSR-VTT dataset.** Models in the second group are additionally pretrained on HowTo100M.

	Text →Video				Video →Text			
	$R@1\uparrow$	$R@5\uparrow$	$R@10\uparrow$	Md$R\downarrow$	$R@1\uparrow$	$R@5\uparrow$	$R@10\uparrow$	Md$R\downarrow$
Random Baseline	0.1	0.5	1.0	500.0	0.1	0.5	1.0	500.0
JSFusion (Yu et al., 2018)	10.2	31.2	43.2	13.0	–	–	–	–
HT100M (Miech et al., 2019)	12.1	35.0	48.0	12.0	–	–	–	–
JPoSE (Wray et al., 2019)	14.3	38.1	53.0	9.0	16.4	41.3	54.4	8.7
CE (Liu et al., 2019)	20.9	48.8	62.4	6.0	20.6	50.3	64.0	5.3
MMT (Gabeur et al., 2020)	24.6	54.0	67.1	4.0	24.4	56.0	67.8	4.0
Ours	**27.4**	**56.3**	**67.7**	**3.0**	**26.6**	**55.1**	**67.5**	**3.0**
VidTranslate (Korbar et al., 2020)	14.7	–	52.8	–	–	–	–	–
HT100M (Miech et al., 2019)	14.9	40.2	52.8	9.0	16.8	41.7	55.1	8.0
NoiseEstimation (Amrani et al., 2020)	17.4	41.6	53.6	8.0	–	–	–	–
UniVL (Luo et al., 2020)	21.2	49.6	63.1	6.0	–	–	–	–
AVLnet (Rouditchenko et al., 2020)	27.1	55.6	66.6	4.0	28.5	54.6	65.2	4.0
MMT (Gabeur et al., 2020)	26.6	57.1	**69.6**	4.0	27.0	57.5	69.7	3.7
Ours-pretrained	**30.1**	**58.5**	69.3	**3.0**	**28.5**	**58.6**	**71.6**	**3.0**

at recall at 1. With pre-training on HowTo100M, the performance further increases across the board. Notably, unlike MMT which uses 7 features, my model uses only 2 features and achieves state-of-the-art in most metrics.

6.2.5 Discussion

In order to better understand the effect of the learning objective, I visualize the soft attention of the best-performing cross-instance reconstruction model in fig. 6.3. As I can see in the top-left square, which shows the pairwise attention between all pairs of videos in the batch, it is highly focused, with the model mostly attending one or two other instances in the batch.

For the first video's caption reconstruction (second row), I find that the model solely attends to another musical performance video that is in the batch, ignoring the others. For the second video (third row), the model focuses on another sample that shows the sea but differs in most other aspects since there are no semantically-equivalent clips in the batch. The third video shares a similar scenario. These examples show that the bottleneck is effective at forcing the model to avoid memorising the video-caption association of each clip in isolation, and attempt to match

Table 6.4: **Retrieval performance on the VATEX dataset**

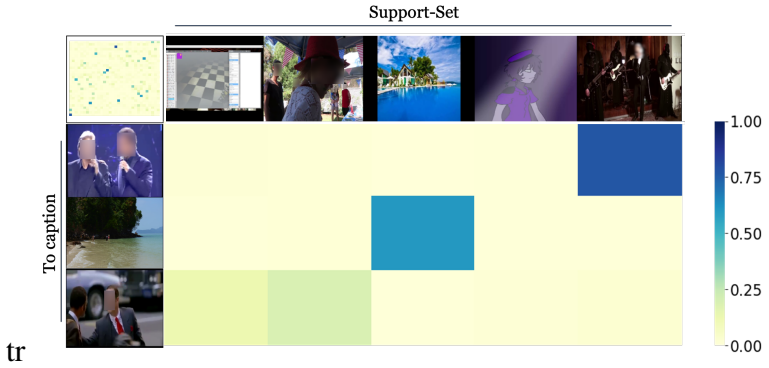
	Text →Video				Video →Text			
	$R@1\uparrow$	$R@5\uparrow$	$R@10\uparrow$	Md$R\downarrow$	$R@1\uparrow$	$R@5\uparrow$	$R@10\uparrow$	Md$R\downarrow$
Random Baseline	0.2	0.7	1.05	2000.5	0.02	0.1	1.02	2100.5
VSE (Kiros et al., 2014)	28.0	64.3	76.9	3.0	–	–	–	–
VSE++ (Faghri et al., 2018)	33.7	70.1	81.0	2.0	–	–	–	–
Dual (He et al., 2016a)	31.1	67.4	78.9	3.0	–	–	–	–
HGR (Chen et al., 2020a)	35.1	73.5	83.5	2.0	–	–	–	–
Ours	**44.6**	**81.8**	**89.5**	**1.0**	**58.1**	**83.8**	**90.9**	**1.0**
Ours-pretrained	**45.9**	**82.4**	**90.4**	**1.0**	**61.2**	**85.2**	**91.8**	**1.0**

Table 6.5: **Retrieval performance on ActivityNet**

	Text →Video				Video →Text			
	$R@1\uparrow$	$R@5\uparrow$	$R@50\uparrow$	Md$R\downarrow$	$R@1\uparrow$	$R@5\uparrow$	$R@50\uparrow$	Md$R\downarrow$
Random Baseline	0.02	0.1	1.02	2458	0.02	0.1	1.02	2458
FSE(Zhang et al., 2018a)	18.2	44.8	89.1	7.0	16.7	43.1	88.4	7.0
CE (Liu et al., 2019)	18.2	47.7	91.4	6.0	17.7	46.6	90.9	6.0
HSE (Zhang et al., 2018a)	20.5	49.3	–	–	18.7	48.1	–	–
MMT (Gabeur et al., 2020)	22.7	54.2	93.2	5.0	22.9	54.8	93.1	4.3
Ours	**26.8**	**58.1**	**93.5**	**3.0**	**25.5**	**57.3**	**93.5**	**3.0**
MMT-pretrained (Gabeur et al., 2020)	28.7	61.4	94.5	3.3	**28.9**	**61.1**	94.3	4.0
Ours-pretrained	**29.2**	**61.6**	**94.7**	**3.0**	28.7	60.8	**94.8**	**2.0**

Table 6.6: **Retrieval performance on the MSVD dataset**

	Text →Video				Video →Text			
	$R@1\uparrow$	$R@5\uparrow$	$R@10\uparrow$	Md$R\downarrow$	$R@1\uparrow$	$R@5\uparrow$	$R@10\uparrow$	Md$R\downarrow$
VSE (Kiros et al., 2014)	12.3	30.1	42.3	14.0	–	–	–	–
VSE++ (Faghri et al., 2018)	15.4	39.6	53.0	9.0	–	–	–	–
Multi. Cues (Mithun et al., 2018a)	20.3	47.8	61.1	6.0	–	–	–	–
CE (Liu et al., 2019)	19.8	49.0	63.8	6.0	–	–	–	–
Ours	**23.0**	**52.8**	**65.8**	**5.0**	**27.3**	**50.7**	**60.8**	**5.0**
Ours-pretrained	**28.4**	**60.0**	**72.9**	**4.0**	**34.7**	**59.9**	**70.0**	**3.0**



tr

Figure 6.3: **Support-set attention map.** Attention scores of all pairs in a batch (top-left square) and a subset of rows/columns (other squares) on VTT.

other clips more broadly, since an exact (or very close) match is not guaranteed.

Summary

In this section, for learning video-text representation at scale, I studied classic contrastive learning methods such as the triplet loss. I suggested that the contrastive approach might pull apart videos and captions even when they are semantically equivalent, which may hinder downstream retrieval performance. To mitigate this effect, I propose to consider a captioning pretext task as an additional learning objective. In particular, I show that cross-instance captioning can encourage the representation to pull together videos that share a similar caption, and are thus likely to be equivalent for retrieval. Leveraging these ideas, my model achieves state-of-the-art performance on the text-to-video and video-to-text retrieval tasks, on three datasets.

While I demonstrated these ideas in the specific case of text-to-video retrieval, they can in principle generalize to any setting that utilizes a contrastive loss, including self-supervised learning, provided that it is possible to learn reasonable conditional generators of a modality or data stream given another.

6.3 Cross-Lingual Transfer of Vision-Language Models

6.3.1 Motivation

One of the key challenges at the intersection of computer vision (CV) and natural language processing (NLP) is building versatile vision-language models that not only work in English, but in all of the world's approximately 6,500 languages. Since collecting and annotating task-specific parallel multimodal data in all languages is impractical, a framework that makes vision-language models generalize across languages is highly desirable.

One technique that has shown promise to greatly improve the applicability of NLP models to new languages is *zero-shot cross-lingual transfer*, where models trained on a source language are applied as-is to a different language without any additional annotated training data (Täckström et al., 2012; Klementiev et al., 2012; Cotterell and Heigold, 2017; Chen et al., 2018a; Neubig and Hu, 2018). In particular, recent techniques for cross-lingual transfer have demonstrated that by performing unsupervised learning of language or translation models on many languages, followed by downstream task fine-tuning using only English annotation, models can nonetheless generalize to a non-English language (Wu and Dredze, 2019; Lample and Conneau, 2019; Huang et al., 2019a; Artetxe et al., 2020; Hu et al., 2020). This success is attributed to the fact that many languages share a considerable amount of underlying vocabulary or structure. At the vocabulary level, languages often have words that stem from the same origin, for instance, "desk" in English and "Tisch" in German both come from the Latin "discus". At the structural level, all languages have a recursive structure, and many share traits of morphology or word order.

For cross-lingual transfer of vision-language models, the visual information is clearly an essential element. To this end, I make an important yet under-explored step to incorporate visual-textual relationships for improving multilingual models (Devlin et al., 2019a; Artetxe et al., 2020). While spoken languages could be different, all humans share similar vision systems, and many visual concepts can be understood universally (Sigurdsson et al., 2020; Zhang et al., 2020a). For example, while  is termed "cat" for an English speaker and "chat" for a French speaker; they understand  similarly. I leverage this observation to learn to associate sentences in different languages with visual concepts for promoting cross-lingual transfer of vision-language models.

In this section, I focus on multilingual text-to-video search tasks and propose a Transformer-based video-text model to learn contextual multilingual multimodal representations. My vanilla model yields state-of-the-art performance in multilingual text→video search when trained with multilingual annotations. However, under the zero-shot setting, rather surprisingly, there is a

significant performance gap between English and non-English queries (see §6.3.6 for details). To resolve this problem, motivated by recent advances in large-scale language model (Artetxe et al., 2020) and multimodal pre-training (Lu et al., 2019; Miech et al., 2019; Patrick et al., 2021a), I propose a multilingual multimodal pre-training (MMP) strategy to exploit the weak supervision from large-scale multilingual text-video data. I construct the Multilingual-HowTo100M dataset, that extends the English HowTo100M (Miech et al., 2019) dataset to contain subtitles in 9 languages for 1.2 million instructional videos.

My method has two important benefits. First, compared to pre-training on English-video data only, pre-training on multilingual text-video data exploits the additional supervision from a variety of languages, and therefore, enhances the search performance on an individual language. Second, by exploiting the visual data as an implicit "pivot" at scale, my methods learns better alignments in the multilingual multimodal embedding space (*e.g.*, "cat"-  -"chat"), which leads to improvement in zero-shot cross-lingual transfer (*e.g.*, from "cat"-  to "chat"- ) of vision-language models. I elaborate the model details in the following sections.

6.3.2 Prior Work

Cross-lingual Transfer Cross-lingual transfer has proven effective in many NLP tasks including dependency parsing (Schuster et al., 2019), named entity recognition (Rahimi et al., 2019), sentiment analysis (Barnes et al., 2019), document classification (Schwenk and Li, 2018), and question answering (Lewis et al., 2020; Artetxe et al., 2020). Recently, XTREME (Hu et al., 2020) was proposed to evaluate the cross-lingual transfer capabilities of multilingual representations across a diverse set of NLP tasks and languages. However, a comprehensive evaluation of multilingual multimodal models on zero-shot cross-lingual transfer capabilities is still missing. To my best knowledge, my work in this thesis is the first attempt to investigate and improve zero-shot cross-lingual transfer of vision-language models.

6.3.3 Multilingual Multimodal Transformers

I consider the problem of learning multilingual multimodal representations from a corpus $\mathcal{C}$ of video-text pairs $\{(x_i, v_i)\}_{i=1}^{C}$, where v_i is a video clip and x_i is its corresponding text (caption or transcription) that is written in one of K languages. My goal is to learn a shared multilingual text encoder $c_x = \Phi(x)$ and a video encoder $c_v = \Psi(v)$, both of which project the input to a shared D-dimensional embedding space $c_v, c_t \in \mathbb{R}^D$, where semantically similar instances (*i.e.*, paired

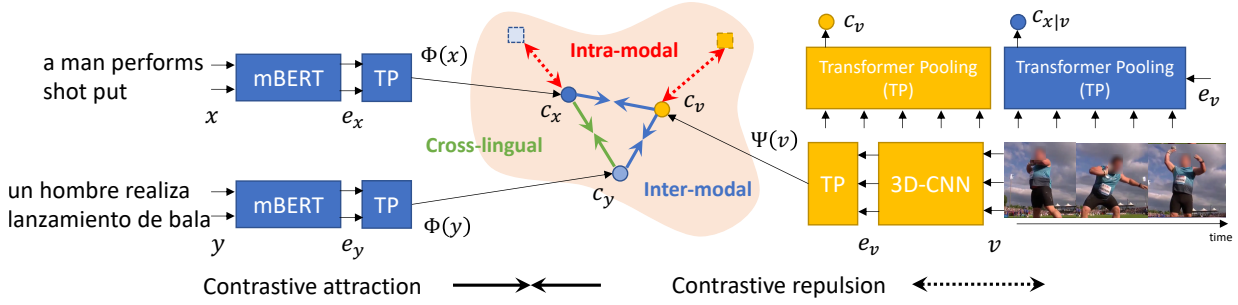


Figure 6.4: An overview of my video-text model for learning contextual multilingual multimodal representations. I utilize *intra-modal*, *inter-modal*, and conditional *cross-lingual* contrastive objectives to align (x, v, y) where x and y are the captions or transcriptions in different languages of a video v. TP: Transformer pooling head.

(x_i, v_i)) are closer to each other than the dissimilar ones (*i.e.*, $(x_i, v_j), i \neq j$). In the following, I denote a batch of multilingual text-video samples as $\mathcal{B} = \{(x_i, v_i)\}_{i=1}^{B}\} \text{ where } \mathcal{B} \subset \mathcal{C}$.

Figure 6.4 gives an overview of the proposed method. My text encoder consists of a multilingual Transformer (*e.g.* multilingual BERT (Devlin et al., 2019a)) and a text Transformer pooling head (explained below). Similarly, my video encoder consists of a 3D-CNN (*e.g.* R(2+1)D network (Tran et al., 2018)) and a video Transformer pooling head. I use these multilingual multimodal Transformers to encode text and video for alignment.

Unlike prior multilingual text-image models (Gella et al., 2017; Kim et al., 2020; Huang et al., 2019b) that utilize word embeddings and RNNs, my multilingual text encoder is built on a multilingual Transformer that generates contextual multilingual representations $e_x \in \mathbb{R}^{N \times D}$ to encode a sentence x containing N words. I employ an additional 2-layer Transformer which I will call a "Transformer pooling head (TP)" as it serves as a pooling function to selectively encode variable-length sentences and aligns them with the corresponding visual content. I use the first output token of the second Transformer layer as the final sentence representation. Precisely, I set $c_x = \text{Trans}_x^{(2)}(\text{query=key=value}=e_x)[0]$ where $\text{Trans}_x^{(2)}$ is a 2-layer stack of Transformers (Vaswani et al., 2017) with e_x as the (query,key,value) in the multi-head attention. Note that I use the same text encoder to encode sentences in all languages.

For encoding videos, my model uses pre-trained 3D-CNNs that encode spatial-temporal context in a video. For a M-second video v, I apply R(2+1)D (Tran et al., 2018) and S3D (Miech et al., 2020) networks to its frames, concatenate network outputs, and apply a linear layer to encode the visual input, $e_v \in \mathbb{R}^{M \times D}$, to the model. Similarly to the text part, I employ a two-layer Transformer as the pooling head to encode videos with different lengths into fixed-

length representations. Formally, I set $c_v = \text{Trans}_v^{(2)}(\text{query=key=value=}e_v)[0]$. Since videos are typically long and have a high frame rate (*e.g.*, 30 fps), it is infeasible to update 3D-CNNs simultaneously and therefore, I use pre-extracted video features. My model is parameterized by $\theta = \theta_{\text{mBERT}} \cup \theta_{\text{Trans}_x} \cup \theta_{\text{Trans}_v}$.

Multilingual Text-Video Alignment For learning multimodal representations, the common practice is to minimize a contrastive objective to map the associated (video, text) embeddings to be near to each other in a shared embedding space. The inter-modal max-margin triplet loss has been widely studied in video-text (Yu et al., 2018; Liu et al., 2019) and image-text (Kim et al., 2020; Burns et al., 2020; Huang et al., 2019b) research. In this work, I generalize and model all *inter-modal*, *intra-modal*, and *cross-lingual* instances with a noise contrastive estimation objective (NCE) (Gutmann and Hyvärinen, 2010; van den Oord et al., 2018; Chen et al., 2020b).

Inter-modal NCE Let $\mathcal{X}$ and $\mathcal{V}$ denote the subsets of the sampled sentences in multiple languages and videos in $\mathcal{B}$, respectively. And let $s(a, b) = \frac{a^T b}{\|a\|\|b\|}$ be the cosine similarity measure. I use an (inter-modal) NCE objective defined as:

$$\mathcal{L}(\mathcal{X}, \mathcal{V}) = -\frac{1}{B}\sum_{i=1}^{B} \log \ell^{\text{NCE}}(\Phi(x_i), \Psi(v_i)), \tag{6.5}$$

where

$$\ell^{\text{NCE}}(c_x, c_v) = \frac{e^{s(c_x, c_v)}}{e^{s(c_x, c_v)} + \sum_{(x', v') \sim \mathcal{N}} e^{s(c_{x'}, c_{v'})}} \tag{6.6}$$

In inter-modal NCE, $\mathcal{L}^{\text{inter}} = \mathcal{L}(\mathcal{X}, \mathcal{V})$, the noise $\mathcal{N}$ is a set of "negative" video-text pairs sampled to enforce the similarity of paired ones are high and and those do not are low. Following Miech et al. (2020), I set the negatives of (x_i, v_i) as other x_j and $v_j, j \neq i$ in $\mathcal{B}$.

Intuitively, inter-modal NCE draws paired (semantically similar) instances closer and pushes apart non-paired (dissimilar) instances. Note that I do not distinguish language types in $\mathcal{X}$ and the sentences in all possible languages will be drawn towards their corresponding videos in the shared multilingual text-video embedding space.

Intra-modal NCE Beyond cross-modality matching, I leverage the intra-modal contrastive objective to learn and preserve the underlying structure within the video and text modality. For example, *Corgi* should be closer to *Husky* than *Balinese*. Prior image-text work (Gella et al., 2017; Huang et al., 2019c) utilizes a triplet loss to maintain such neighborhood relationships. Inspired by recent success in self-supervised image and video representation learning (Yalniz et al., 2019; Ghadiyaram et al., 2019), my model leverages intra-modal NCE that constrains the

learned representations to be invariant against noise and to maintain the within-modality structure simultaneously. I minimize the following intra-modal NCE loss:

$$\mathcal{L}^{\text{intra}} = \mathcal{L}(\mathcal{X}, \mathcal{X}^m) + \mathcal{L}(\mathcal{V}, \mathcal{V}^m), \tag{6.7}$$

where $\mathcal{X}^m$ and $\mathcal{V}^m$ are the noised version of the original sentences and videos. For noising, I randomly mask 5% of the multilingual text tokens and video clips. I optimize my model by

$$\min_{\theta} \mathcal{L}^{\text{inter}} + \mathcal{L}^{\text{intra}} \tag{6.8}$$

When Visually-Pivoted Multilingual Annotations Are Available In many multilingual multimodal datasets, there are sentences in different languages that describe a shared visual context. For example, 10 English and 10 Chinese descriptions are available for each video in VATEX. With these visually-pivoted (weakly paralleled) sentences (x, y), I further revise the contrastive objectives to leverage this additional supervisory signal. Given a visually-pivoted corpus $\mathcal{C}^p$ that contains all possible combination of visually-pivoted pairs $\{(x_i, v_i, y_i)\}_{i=0}^{C_p}$, I sample batches $\mathcal{B}^p = \{(x_i, v_i, y_i)\}_{i=1}^{\mathcal{B}^p}, \mathcal{B}^p \subset \mathcal{C}^p$ and revise the contrastive objective as:

$$\mathcal{L}^{\text{inter}} = \mathcal{L}(\mathcal{X}, \mathcal{V}) + \mathcal{L}(\mathcal{Y}, \mathcal{V}) \tag{6.9}$$

$$\mathcal{L}^{\text{intra}} = \mathcal{L}(\mathcal{X}, \mathcal{X}^m) + \mathcal{L}(\mathcal{Y}, \mathcal{Y}^m) + \mathcal{L}(\mathcal{V}, \mathcal{V}^m) \tag{6.10}$$

Visual-pivoted Cross-lingual NCE Inspired by Translation Language Modeling (TLM) in XLM (Lample and Conneau, 2019), I propose a multimodal TLM-like contrastive objective which promotes alignments of descriptions in different languages that describe the same video. I use the intuition that conditioned on a video, the descriptions (need not to be translation pairs) in different languages would likely be semantically similar. To this end, I set the cross-lingual NCE as:

$$\mathcal{L}^{\text{cross}} = \mathcal{L}(\mathcal{X}|\mathcal{V}, \mathcal{Y}|\mathcal{V}) \tag{6.11}$$

For visually-pivoted sentences, as shown in Figure 6.4, I generate their representations conditioned on the video they describe. I extend the *key* and *value* of multihead attention with the additional visual content e_v and generate new $c_{x|v}$ and $c_{y|v}$ for matching. Specifically, my model employs $c_{x|v} = \text{Trans}_x^{(2)}(\text{query}{=}e_x, \text{key}{=}\text{value}{=}e_x||e_v)[0]$. With the access to (visually-pivoted) multilingual annotations, I optimize my model by

$$\min_{\theta} \mathcal{L}^{\text{inter}} + \mathcal{L}^{\text{intra}} + \mathcal{L}^{\text{cross}} \tag{6.12}$$

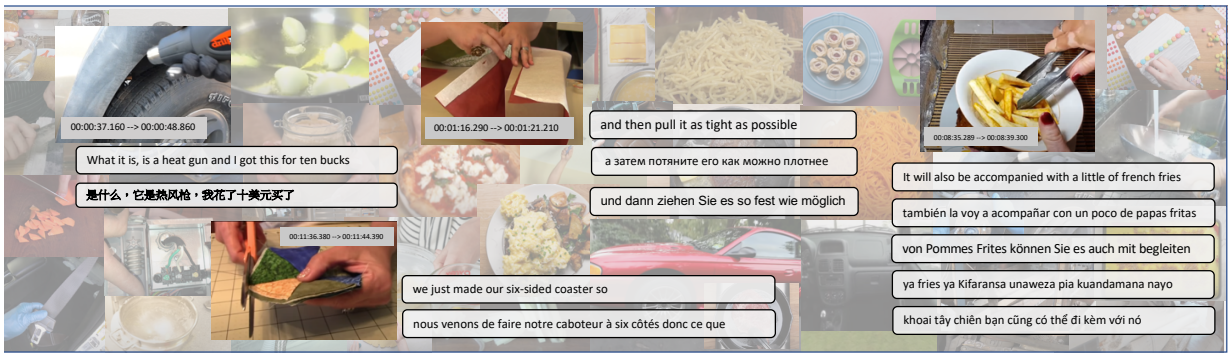


Figure 6.5: Video clips and the corresponding multilingual subtitles in Multi-HowTo100M.

At the inference time, I simply apply $c_x = \Phi(x)$ and $c_v = \Psi(v)$ to encode multilingual text queries and videos. For text-to-video search, I sort videos according to their cosine similarity scores to the text query.

6.3.4 The Multilingual HowTo100M Dataset

As large-scale pre-training has been shown important in recent NLP and vision-language models, I construct the **Multilingual HowTo100M** dataset (Multi-HowTo100M) to facilitate research in multilingual multimodal learning. The original HowTo100M (Miech et al., 2019) dataset is a large-scale video collection of 1.2 million instructional videos (around 138 million clips/segments) on YouTube, along with their automatic speech recognition (ASR) transcriptions as the subtitles. For each video in HowTo100M, I crawl and collect the multilingual subtitles provided by YouTube, which either consist of user-generated subtitles or those generated by Google ASR and Translate in the absence of user-generated ones. Essentially, I collect video subtitles in 9 languages: English (*en*), German (*de*), French (*fr*), Russian (*ru*), Spanish (*es*), Czech (*cz*), Swahili (*sw*), Chinese (*zh*), Vietnamese (*vi*).

At the time of dataset collection (May 2020), there are 1.1 million videos available, each with subtitles in 7-9 languages. The video length ranges from 1 minute to more than 20 minutes. I utilize Multi-HowTo100M for multilingual multimodal pre-training to exploit the weak supervision from large-scale multilingual text-video data. In Figure 6.5, I provide a visualization of few instances sampled in Multi-HowTo100M with the corresponding video frame, timestamp, and transcriptions in different languages. Please refer to Appendix for more details and dataset statistics.

6.3.5 Empirical Evaluation

In this section, I first describe my experimental setup (§6.3.5-6.3.5). In §6.3.5, I conduct ablation studies to validate the effectiveness of proposed multilingual text-video model . With the best models at hand, I investigate their zero-shot cross-lingual transferability in §6.3.6, where I showcase that the proposed multilingual multimodal pre-training serves as the key facilitator. I then verify the superior text→video search performance of the proposed method under the monolingual, multilingual, and cross-modality settings in §6.3.7.

Evaluation Datasets

MSR-VTT (VTT) (Xu et al., 2016) contains 10K videos, where each video is annotated with 20 captions. Additionally, I created pseudo-multilingual data by translating the English captions into 8 languages with off-the-shelf machine translation models.[1] I use the official training set (6.5K videos) and validation set (497 videos). I follow the protocol in Miech et al. (2019); Liu et al. (2019) which evaluates on text→video search with the 1K testing set defined by Yu et al. (2018).
VATEX (Wang et al., 2019) is a multilingual (Chinese and English) video-text dataset with 35K videos. Five (*en*,*zh*) translation pairs and five non-paired *en* and *zh* descriptions are available for each video. I use the official training split (26K videos) and follow the testing protocol in Chen et al. (2020a) to split the validation set equally into 1.5K validation and 1.5K testing videos.
Multi30K (Elliott et al., 2016) is a multilingual extension of Flickr30K (Young et al., 2014). For each image, there are two types of annotations available: (1) One parallel (English,German,French,Czech) translation pair and (2) five English and five German descriptions collected independently. The training, validation, and testing splits contain 29K, 1K, and 1K images respectively.

Implementation Details

For the video backbone, I use a 34-layer, R(2+1)-D (Tran et al., 2018) network pre-trained on IG65M (Ghadiyaram et al., 2019) and a S3D (Miech et al., 2020) network pre-trained on HowTo100M. I pre-extract video features and concatenate the two 3D-CNN outputs to form $e_x \in \mathbb{R}^{M\times 1024}$ as a video input.

For the text backbone, I use multilingual BERT (mBERT) (Devlin et al., 2019a) or XLM-Roberta-large (XLM-R) (Artetxe et al., 2020), where the latter achieves near SoTA zero-shot cross-lingual transfer performance for NLP tasks. Following Hu et al. (2020), instead of using

[1] https://marian-nmt.github.io/

Text-B	Video-B	R@1↑	R@5↑	R@10↑
XLM-R	S3D	19.5	49.0	62.8
XLM-R	R(2+1)D	19.0	49.5	63.2
XLM-R	R+S	**21.0**	**50.6**	**63.6**
mBERT	R+S	19.9	49.8	62.5

Table 6.7: **Text and Video (B)ackbone comparison.**

the top layer, I output the 12-th layer in XLM-R and mBERT. For vision-language tasks, I freeze layers below 9 as this setup empirically performs the best.

My model employs a 2-layer Transformer with 4-head attention for the text and video transformer pooling (TP) modules. The embedding dimension D is set to 1024. I use the Adam (Kingma and Ba, 2014) optimizer and a 0.0002 learning rate to train my model for 16 (pre-training) and 10 (fine-tuning) epochs. The softmax temperature in all noise contrastive objectives is set to 0.1.

Experimental Setup

I use Multi-HowTo100M for multilingual multimodal pre-training (MMP). For each video, I randomly sample the start and end time to construct a video clip. For a video clip, I randomly sample one language type each time from 9 languages and use the consecutive ASR transcriptions that are closest in time to compose (text-video) pairs for training. For simplicity and speed purposes, I follow the training protocol of XLM-R to pre-train on a multilingual corpus *wihtout* using translation pairs, *i.e.*, I use multilingual text-video pairs (x, v) but no translation pairs from Multi-HowTo100M and utilize only inter- and intra-modal NCE (Eq. 6.5-6.7) for MMP.

I fine-tune the model on VTT, VATEX, and Multi30K to evaluate on text→video search tasks. In the zero-shot cross-lingual transfer experiments, I use only English-video data and fine-tune with Eq. 6.5-6.7. I then test the model with non-English queries. When annotations in additional languages are available (by humans in VATEX and Multi30K; by MT models (*i.e.*, *translate-train*) in VTT), I utilize all available multilingual annotations (*i.e.*, fully supervised) and iterate over all possible (x, v, y) pairs to train with Eq. 6.9-6.11 to demonstrate the strong performance target for evaluating zero-shot cross-lingual transfer on VTT and to compare fairly with other fully-supervised baselines in multilingual text→video search on VATEX and Multi30K. I report the standard recall at k (R@k) metrics (higher is better).

T layers	V layers	R@1↑	R@5↑	R@10↑
1	1	20.0	50.3	63.2
2	1	20.1	50.5	63.8
2	2	**21.0**	**50.6**	63.6
2*	2*	20.7	50.5	63.3
4	4	20.8	50.4	**63.8**

Table 6.8: **Architecture comparison.** Number of multilingual multimodal transformer layers. *:Weight sharing between video and text transformers.

Objective	Inter	Intra	Cross	R@1↑	R@5↑	R@10↑
Triplet	✓			13.3	36.0	55.2
Triplet	✓	✓		20.9	49.3	63.0
NCE	✓			21.4	49.3	61.1
NCE	✓	✓		21.0	50.6	63.6
NCE*	✓	✓		21.3	50.7	63.5
NCE*	✓	✓	✓	**21.5**	**51.0**	**63.8**

Table 6.9: **Objective comparison.** *Training with additional machine translated *de*-video and *fr*-video pairs.

Comparison Experiments and Ablations

In this section, I ablate and compare different text/video encoders, Transformer model architectures, and learning objectives for English→video search on VTT.

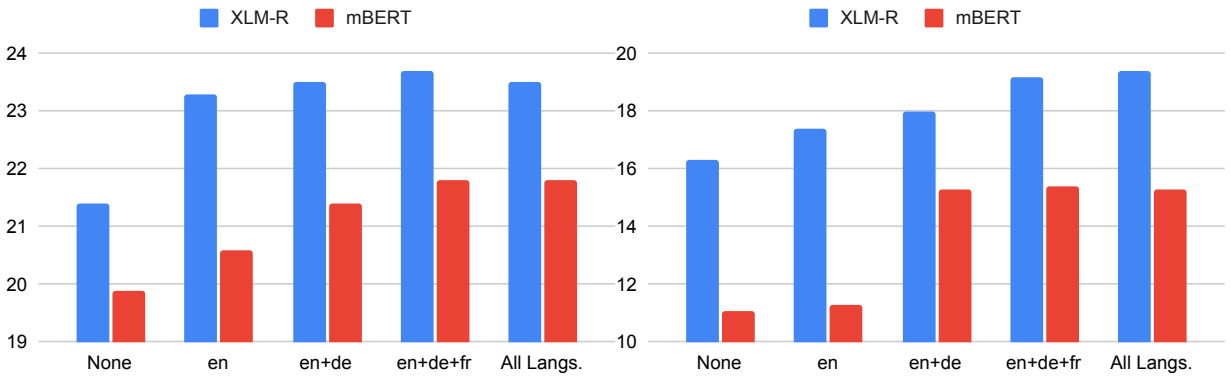
Text and Video Encoders. Table 6.7 compares different text and video encoder backbones. For the visual encoders, while R(2+1)D outperforms S3D, the simple concatenation (*i.e.*, early-fusion) of their output features provides a $1.5 \sim 2.0$ improvement in R@1. For the text encoder, XLM-R significantly outperforms mBERT.

Transformer Pooling. Table 6.8 compares various configurations of the proposed Transformer pooling module. I observe that a simple 2-layer Transformer achieves the best performance. Weight sharing of the video and text Transformer slightly degrades the performance. Therefore, I choose to separate them.

Learning Objective. From Table 6.9, the intra-modal contrastive objective is important for both NCE and Triplet loss. In general, the NCE loss outperforms the Triplet loss. The proposed inter-modal and intra-modal NCE objective achieves the best performance. When captions in multiple

Model	*en*	*de*	*fr*	*cs*	*zh*	*ru*	*vi*	*sw*	*es*	Avg↑
mBERT	19.9	11.1	11.6	8.2	6.9	7.9	2.7	1.4	12.0	9.1
mBERT-MP	20.6	11.3	11.9	8.0	7.1	7.7	2.5	1.1	12.5	9.2
mBERT-MMP	21.8	15.0	15.8	11.2	8.4	11.0	3.7	3.4	15.1	11.7
XLM-R	21.0	16.3	17.4	16.0	14.9	15.4	7.7	5.7	17.3	14.7
XLM-R-MP	23.3	17.4	18.5	17.1	16.3	17.0	8.1	6.2	18.5	15.8
XLM-R-MMP	**23.8**	**19.4**	**20.7**	**19.3**	**18.2**	**19.1**	**8.2**	**8.4**	**20.4**	**17.5**
mBERT + translated VTT	19.6	18.2	18.0	16.9	16.2	16.5	8.4	13.0	18.5	16.1
mBERT-MMP + translated VTT	21.5	19.1	19.8	18.3	17.3	18.3	8.9	14.1	20.0	17.4
XLM-R + translated VTT	21.5	19.6	20.1	19.3	18.9	19.1	10.3	12.5	18.9	17.8
XLM-R-MMP + translated VTT	**23.1**	**21.1**	**21.8**	**20.7**	**20.0**	**20.5**	**10.9**	**14.4**	**21.9**	**19.4**

Table 6.10: **Recall@1 of multilingual text→video search on VTT.** Upper: Zero-shot cross-lingual transfer. Lower: Performance with synthesized pseudo-multilingual annotations for training. MMP: multilingual multimodal pre-training on Multi-HowTo100M. MP: Multimodal (English-Video) pre-training on HowTo100M.



(a) English→Video (b) Zero-shot German→Video

Figure 6.6: R@1 trends in languages used for multilingual multimodal pre-training. Left: English→video search. Right: Zero-shot German→video search.

languages are available, cross-lingual NCE additionally provides a consistent improvement.

6.3.6 Zero-Shot Cross-Lingual Transfer

Table 6.10 shows the multilingual text→video search results on VTT. With the best English-video models at hand (with either mBERT or XLM-R as the text backbone), I first investigate how well

Figure 6.7: Qualitative multilingual (*en*, *ru*, *vi*, *zh*) text→video search results on VTT.

these models transfer to other non-English languages under the zero-shot setting. I then analyze the benefit of the proposed multilingual multimodal pre-training.

The upper section shows the zero-shot results. Unlike cross-lingual transfer in NLP tasks, employing multilingual Transformers in vision-language tasks apparently does not generalize well across languages. For example, there is a significant drop in R@1 (19.9→11.1 (-44%) with mBERT, 21.0→16.3 (-24%) with XLM-R) when directly applying English-finetuned model to German→video search. For comparison, there is only a -10% degradation for XLM-R on $en \rightarrow de$ cross-lingual transfer in XNLI (Conneau et al., 2018b). Multimodal (English-video) pre-training (MP) on HowTo100M only improves average R@1 (+0.1 or mBERT and +1.1 for XLM-R) compared to model-from-scratch. In contrast, the proposed multilingual multimodal pre-training (MMP) is shown to be the key facilitator for zero-shot cross-lingual transfer. MMP improves German→Video search (11.1→15.0, +35% for mBERT, and 16.3→19.4, +20% for XLM-R) and achieves $2.6 \sim 2.8$ improvement in average R@1. I attribute the effectiveness of MMP to learning improved alignments between multilingual textual and visual context in the shared embedding space, as relatively balanced improvements between English→video and non-English→video is observed with fine-tuning.

Figure 6.6 demonstrates the trend of R@1 while incrementally incorporating additional languages for MMP. For XLM-R, the improvement in R@1 asymptotically converges when pre-training with more multilingual text-video pairs. On the other hand, for zero-shot German→video search, pre-training with more languages keeps improving the search performance, even though the additional language (*e.g.*, French) is different from the target language (*i.e.*, German).

The lower section of Table 6.10 shows the results of models fine-tuned with (synthesized) pseudo-multilingual annotations. It can be regarded as the *translate-train* scenario, which serves as a strong performance target for evaluating zero-shot cross-lingual transfer, as discussed in (Lample and Conneau, 2019; Hu et al., 2020). Both mBERT and XLM-R yield better performance across non-English languages with the in-domain translated pseudo-multilingual annotations. However,

for English→video search, a 0.7 degradation is observed compared to the zero-shot setting. It is likely due to the noise in the translated captions. Notably, there is still a performance gap between zero-shot and translate-train settings for models with mBERT. In contrast, the gap is much smaller for models with XLM-R. In the following sections, I refer `Ours-MMP` as my best model with XLM-R as the text backbone and compare it with other state-of-the-art methods.

Qualitative Results

Figure 6.7 shows the multilingual text→video search results with Ours-MMP (VTT:*en*-only) on VTT under the zero-shot setup. Note that only one shared English-finetuned model is used for text→video search in all languages. As demonstrated, the proposed model successfully retrieves the correct videos with English (*en*) and Russian (*ru*) queries. The other top-ranked videos also share similar visual appearance to the correct one. For zero-shot transferring of the English-finetuned model to distant languages such as Vietnamese (*vi*) and Chinese (*zh*), I observe that there is still limitation for the zero-shot models to understand abstract concepts (*e.g.*, "space project") and associate small objects (*e.g.*, "microphone") with the text queries in distant languages.

6.3.7 Comparison to Supervised State of the Art

Model	R@1↑	R@5↑	R@10↑
JSFusion (Yu et al., 2018)	10.2	31.2	43.2
JPoSE (Wray et al., 2019)	14.3	38.1	53.0
VidTrans† (Korbar et al., 2020)	14.7	–	52.8
HT100M† (Miech et al., 2019)	14.9	40.2	52.8
Noise† (Amrani et al., 2020)	17.4	41.6	53.6
CE^2 (Liu et al., 2019)	20.9	48.8	62.4
Ours(VTT:en-only)	21.0	50.6	63.6
Ours-MMP (VTT:en-only)	**23.8**	**52.6**	**65.0**

Table 6.11: English→video search performance on VTT. †: Models with pre-training on HowTo100M.

Model	English to Video			Chinese to Video		
	R@1↑	R@5↑	R10↑	R@1↑	R@5↑	R@10↑
VSE (Kiros et al., 2014)	28.0	64.3	76.9	-	-	-
VSE++ (Faghri et al., 2018)	33.7	70.1	81.0	-	-	-
Dual (He et al., 2016a)	31.1	67.4	78.9	-	-	-
HGR (Chen et al., 2020a)	35.1	73.5	83.5	-	-	-
Ours (VATEX:en-only)	43.5	79.8	88.1	23.9	55.1	67.8
Ours-MMP (VATEX:en-only)	**44.4**	80.5	88.7	29.7	63.2	75.5
Ours-MMP (VATEX:en, zh)	44.3	**80.7**	**88.9**	**40.5**	**76.4**	**85.9**

Table 6.12: Multilingual text→video search on VATEX.

English→Video Search on VTT

Table 6.11 shows the comparison of English→video models on VTT. For a fair comparison to other baselines, my model fine-tunes only with the original English annotations on VTT. The results show that my model outperforms other baselines by a large margin. Specifically, my model achieves 8.9 R@1 improvement over the original HowTo100M model (Miech et al., 2019) and other recent baselines with pre-training on HowTo100M. Using a smaller set of visual features and training on a smaller (6,513 vs 9,000) training set[2], my model also outperforms CE (Liu et al., 2019) with or without pre-training.

Multilingual Text→Video Search on VATEX

Table 6.12 summarizes English→video and Chinese→video search performance on the VATEX dataset. As observed, my model generalizes well across distant languages (English and Chinese). Under the zero-shot setting where I train with only English-video pairs, my model already outperforms other baselines. However, a clear performance gap between English→video and Chinese→video search is observed, indicating that cross-lingual transfer to a distant language remains challenging even with XLM-R. With the proposed MMP, the gap is significantly closed by 5.8/8.1/7.7 in R@1/5/10. When in-domain human-annotated Chinese captions are available, the performance of my model can further be improved for both languages and it yields new state-of-the-art performance.

[2]CE uses 9,000 videos (VTT training and part of exclusive testing set) for training, while other baselines and my model in Table 6.11 are trained on the official VTT training set which contains 6,513 videos.

Model	M30K # lang.	English to Image R@1↑	R@5↑	R10↑	German to Image R@1↑	R@5↑	R@10↑	Czech to Image R@1↑	R@5↑	R@10↑
OE (Vendrov et al., 2015)	2	21.0	48.5	60.4	25.8	56.5	67.8	-	-	-
VSE++ (Faghri et al., 2018)	2	31.3	62.2	70.9	39.6	69.1	79.8	-	-	-
Pivot (Gella et al., 2017)	2	22.5	49.3	61.7	26.2	56.4	68.4	-	-	-
FB-NMT (Huang et al., 2020a)	2	47.3	75.4	83.5	37.0	64.0	73.1	-	-	-
MULE (Kim et al., 2020)	4	42.2	72.2	81.8	35.1	64.6	75.3	37.5	64.6	74.8
SMALR (Burns et al., 2020)	10	41.8	72.4	82.1	36.9	65.4	75.4	36.7	68.0	78.2
MHA-D (Huang et al., 2019b)	2	50.1	78.1	85.7	40.3	70.1	79.0	-	-	-
Ours (M30K:en-only)	1	48.4	78.3	85.9	31.4	61.1	72.6	33.2	65.2	76.1
Ours-MMP (M30K:en-only)	1	50.0	79.2	86.8	33.8	63.3	74.7	37.9	68.8	78.2
Ours-MMP (M30K:en, de, cs, fr)	4	**51.6**	**80.1**	**87.3**	**45.1**	**75.6**	**85.0**	**46.6**	**75.9**	**83.4**

Table 6.13: Multilingual text→image search on Multi30K. MMP: Multilingual multimodal pre-training.

Cross-Modality Transfer to Multi30K: From Video-Text to Image-Text

To extend my study on zero-shot cross-lingual transfer for image-text tasks, I investigate the feasibility of transferring my video-text model across modalities. I replace the 3D-CNN in the original video-text model with a 2D-CNN to encode the image. In practice, following MHA-D (Huang et al., 2019b), I utilize the Faster-RCNN (Ren et al., 2015) pre-trained in Visual Genome (Krishna et al., 2017b) to extract regional visual features. Essentially, an image is encoded as $e_v = \mathbb{R}^{M \times H}$ where $M = 36$ is the maximum number of visual objects in an image. For models with MMP, I initialize their weights with the model pre-trained on Multi-HowTo100M. To tackle the feature mismatch between 2D-CNN and 3D-CNN, I leverage a linear layer with a doubled learning rate to map 2D-CNN features to the same dimension as 3D-CNN features.

Table 6.13 shows the results on Multi30K. For zero-shot cross-lingual transfer, when trained from scratch (M30K:en-only), my model achieves comparable performance to MHA-D but lags in German→image search since it only uses English annotations. In Ours-MMP, pre-training improves all recall metrics even with modality gap. The average R@1 improvement is 3.2. A larger gain for (relatively) low-resource language such as Czech is observed. Without using any Czech annotations, my zero-shot model with MMP achieves comparable Czech→image search performance to SMALR (Burns et al., 2020), which uses 10 languages including Czech. However, when transferring across modalities and using only English annotations, there are performance gaps between English→Image and German/Czech→Image search, implying that transferring models across modalities is feasible but remains challenging. I consider zero-shot cross-modal

cross-lingual transfer as the future work.

For a fair comparison with other baselines, when trained with annotations in all 4 languages provided by Multi30K, my model greatly outperforms all baselines by large margins in multilingual text→image search. Notably, English→image search performance can further be improved by training with multilingual annotations.

6.4 Summary

In this chapter, I presented methods for large-scale pre-training of multilingual text-video representations. I firstly targeted analyzing the bottlenecks of contrastive video-text representation and provide a generative captioning objective to encourage sharing of video semantics to alleviate the bottlenecks in contrastive learning that could be too strict for video-text. This simple idea ensures that representations are not overly specialized to individual samples, are reusable across the dataset, and results in representations that explicitly encode semantics shared between samples, unlike noise contrastive learning. The proposed method outperforms others by a large margin on MSR-VTT, VATEX, ActivityNet, and MSVD for video-to-text and text-to-video retrieval.

I then constructed multilingual multimodal transformers to additionally model videos with multilingual captions and propose a pre-training strategy on my newly collected Multi-HowTo100M dataset to learn contextual multilingual multimodal representations and to promote cross-lingual generalizations of vision-language models. The results have convincingly demonstrated that multilingual multimodal pre-training is an essential ingredient for zero-shot cross-lingual transfer of vision-language models.

For future work, there are many remaining challenges, such as resolving the performance gap between zero-shot and training with in-domain non-English annotations; as well as techniques to transfer varieties of vision-language models (*e.g.*, VQA (Goyal et al., 2017), TVQA (Lei et al., 2020)) or visually-enhanced NLP models such as unsupervised multimodal machine translation (Huang et al., 2020b). I believe the proposed methodology, and the corresponding resources I released, will spur more research in this direction.

Chapter 7

Multilingual Multimodal Fine-tuning under Limited Supervision

7.1 Overview

In the previous Chapter §6, I have presented my methods for task-agnostic pre-training multilingual multimodal representations with large-scale uncurated data (videos and the corresponding (multilingual) captions) and studied zero-shot cross-lingual transfer of vision-language models. In this chapter, I address the challenges of multilingual multimodal fine-tuning on the end task. In the real-world scenario, English annotations in a multimodal dataset may be scarce due to data privacy or the cost of hiring annotators. Besides, most of the multimodal datasets are English-centered and not multilingual. Developing robust methods that learn under limited English and non-English supervision is therefore critical in multilingual multimodal fine-tuning.

As depicted in the roadmap (Fig. 1.1), in this chapter, I firstly develop a method to alleviate the side-effects of lacking English-vision annotations in cross-modal retrieval (section §7.2). The proposed adversarial attentive alignment model (A3VSE) leverages visual semantics and adversarial learning to alleviate the introduced heterogeneous domain gaps. It improves the robustness when learning under sparse English-vision annotations.

Meanwhile, to enable multilingual multimodal fine-tuning with English-only annotations, which is the common case for most vision-language datasets, I propose to use multimodal machine translation (MT) models to translate training data into non-English languages for multilingual multimodal fine-tuning. As illustrated in the thesis roadmap, I build two types of multimodal MT models: supervised multimodal MT and unsupervised multimodal MT.

In section §7.3, I investigate various approaches to incorporate visual content for MT and show that the visual information, which is available in vision-language tasks and datasets, is crucial to improve text-only MT. Albeit it could feasible to use off-the-shelf multimodal MT models or a parallel corpus to train such MT model for translate-train, in the real-world scenario, such a model or a parallel corpus for training MT models may be unavailable for some languages. To address this issue, in section §7.4 I introduce my unsupervised multimodal MT module that relies on *No* parallel corpora for training the MT model.

The content in this chapter appears in:

1. "Annotation Efficient Cross-Modal Retrieval with Adversarial Attentive Alignment," Po-Yao Huang, Guoliang Kang, Wenhe Liu, Xiaojun Chang, Alexander G Hauptmann, *ACM MM 2019*.
2. "Attention-based Multimodal Neural Machine Translation," Po-Yao Huang, Frederick Liu, Sz-Rung Shiang, Jean Oh, Chris Dyer, *WMT 2016*.
3. "Unsupervised Multimodal Neural Machine Translation with Pseudo Visual Pivoting," Po-Yao Huang, Junjie Hu, Xiaojun Chang, Alexander Hauptmann, *ACL 2020*.

7.2 Learning Multimodal Representations with Less Labels

7.2.1 Motivation

Recently, deep neural networks have made significant advancements for learning joint visual-semantic embeddings (VSE) (Karpathy and Fei-Fei, 2015; Zheng et al., 2017; Huang et al., 2019d). Such success largely attributes to the availability of large-scale, high-quality human-annotated parallel corpora such as the MS-COCO (Lin et al., 2014) and Flickr30K (Young et al., 2014) datasets. Necessarily, there are more than 610,000 and 150,000 annotated image-text pairs in MS-COCO and Flickr30K, respectively. Although models trained with an affluent amount of well-annotated multimodal pairs can achieve reasonable performance, such assumption does not apply to the real-world scenario. For instance, forensic or medical image data D'souza Rhett et al. (2020) are usually very limited and with a strict privacy-control protocol. Also, hiring professionals to label class or verbally describe these data could be prohibitively costly. To this end, developing methods that could learn with minimum supervision is critical to ensure the robustness of these models.

To understand the impact of learning with limited supervision, I analyze cross-modal retrieval

models and observe that these models struggle when only a limited amount of parallel annotations are available. As shown in Fig. 7.1, recent VSE models (Vendrov et al., 2015; Faghri et al., 2018; Nam et al., 2017; Lee et al., 2018) all suffer severe degeneration as the annotations become sparsely available.

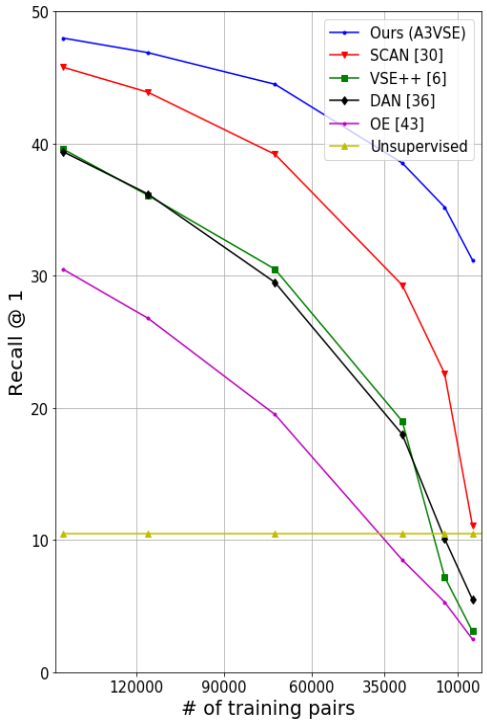


Figure 7.1: Performance degeneration of state-of-the-art cross-modal retrieval models in the text-to-image retrieval task on Flick30K when learning with limited image-text supervision.

The **sparse parallel corpus** (Figure 7.2) is a practical scenario where a large collection of visual data is available but only a small amount of them are annotated with corresponding text descriptions. As annotations become more sparsely available, most recent VSE model experience a severe performance degradation (see Fig. 7.1). To this end, I pose a challenging yet rewarding question: *Can I learn satisfactory visual-semantic embedding with a sparse parallel corpus?* Despite some recent progress (Kiros et al., 2014; Mithun et al., 2018b), learning with small a amount of parallel data is still challenging and to be developed in urgent need.

A straightforward way to deal with a sparse parallel corpus is to utilize the machine-generated semantics of the images directly. In (Mithun et al., 2018b), Mithun *et al.* proposed a webly approach to utilize the global tags of the images. However, without handling the inevitable domain gap between the natural language description and the machine-generated tags properly,

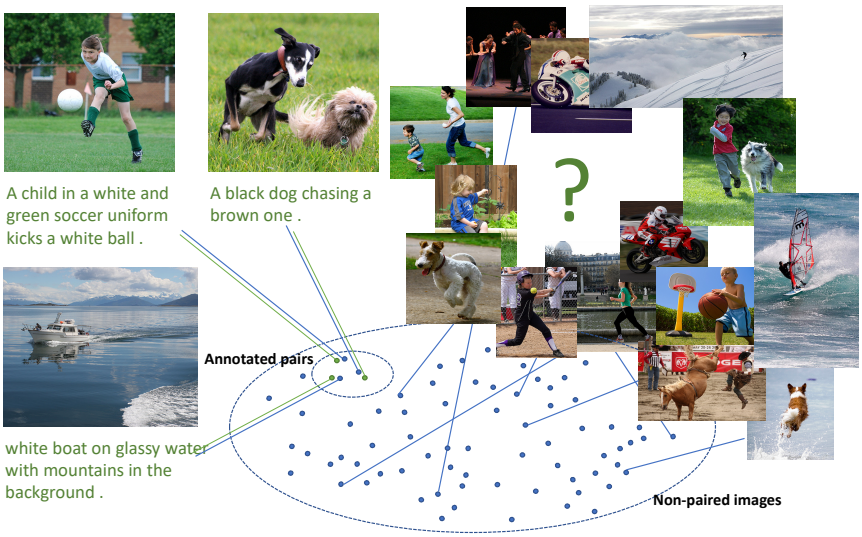


Figure 7.2: A sparsely annotated parallel corpus with abundant un-annotated images and limited (image, natural language sentence) pairs.

the visual-semantic embedding learning could be negatively affected, which largely limits the performance.

To circumvent these issues, inspired by the observation in (Anderson et al., 2018) where bottom-up attention over regional objects aligns well with human's visual system, I propose to utilize "*regional semantics*" which correspond to the regions-of-interest in the un-annotated images and leverage the textual sequences of them to form "pseudo" image-text pairs as the additional weak supervision to conquer the sparsity of image-text annotation. Each regional semantic consists of the category of visual object and its attributes (*e.g.*, *white cat*) which can be automatically extracted with object detection modules (Ren et al., 2015; Abdulla, 2017). With the inferred regional semantics, I develop a novel method to learn the joint visual-semantic embedding space from both the annotated pairs and the inferred pairs efficiently. To minimize the inherent domain gaps between the annotated and un-annotated portion of visual and textual domains, I further impose an attentive alignment with adversarial learning objectives to selectively improve the correlation of semantically close components.

In the following, I first review the prior research and identify the unique contribution by this thesis. I then formally define the problem setup and present the proposed adversarial attentive alignment for learning VSE with synthesis image-text pairs. Finally, I showcase the empirical evaluation and discuss the results.

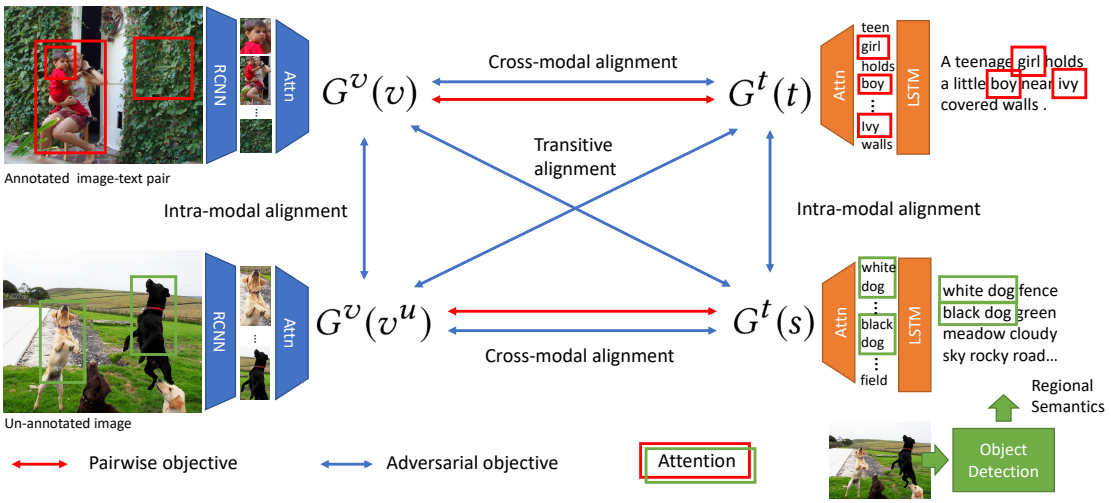


Figure 7.3: The proposed adversarial attentive alignment model (**A3VSE**) for sparsely annotated multimodal corpora. My model incorporates pseudo "image-text" pairs (illustrated as the bottom image-semantic pair) from the sequence of regional semantics of salient visual objects in un-annotated images. The triplet objectives (colored in red) and adversarial objectives (colored in blue) attend and align semantically correlated instances in the joint embedding space while closing the heterogeneous domain gaps between the annotated/un-annotated portion of visual and textual inputs.

7.2.2 Prior Work

Cross-Modal Retrieval with Limited Supervision. Jiang et al. (2018) propose a coupled dictionary learning method to learn the class prototypes that utilize the discriminative information of visual space to improve the less discriminative semantic space. Tsai et al. (2017) augment a typical supervised formulation with unsupervised techniques for learning joint embeddings of visual and textual data. To the best of my knowledge, the most relevant work to my work are (Gong et al., 2014), where the authors resource meta data and image tags (*i.e.*, global semantics) to improve learning of the visual-semantic embedding space. The work in this thesis complements prior work in two perspectives: First, I explore the feasibility of automatic regional semantics as they are more similar to natural language descriptions and leverage them for training improved sequential text encoder. Furthermore, I consider to close the inherent heterogeneous domain gaps with adversarial attentive alignment.

7.2.3 Problem Formulation

I consider a typical scenario where annotated image-text pairs are sparsely available, and un-annotated images are abundant. Let $\mathcal{D}^l = \{I_1, \cdots, I_{N_l}\}$ be an annotated collection of instances where each instance $I_i = (v, t)$ consists of the image v and the corresponding natural language description t. Let $\mathcal{D}^u = \{v^u_1, \cdots, v^u_{N_u}\}$ denotes the collected but un-annotated images. I name $\mathcal{D} = \mathcal{D}^l \cup \mathcal{D}^u$ where $N_l << N_u$, as a **sparse parallel corpus**. I aim to utilize the un-annotated data $\mathcal{D}^u$, together with the annotated data $\mathcal{D}^l$, to learn better visual-semantic embeddings.

7.2.4 Adversarial Attentive Alignment for Improving VSE

Fig. 7.3 illustrates the proposed adversarial attentive alignment model for learning visual-semantic embeddings (**A3VSE**). The proposed model jointly leverages the strong supervision from the annotated image-text pairs and the weak supervision from the inferred image-semantics pairs. Furthermore, A3VSE employs attentive adversarial objectives to selectively align entities from the annotated and un-annotated portion of visual and textual inputs and narrow the domain gaps in between.

For encoding context in each modality, I apply an attention network that focuses on specific encoded regions/ tokens of inputs with respect to the global context from the same modality. I leverage a K-head context-aware attention network to capture the interactions between encoded entities and select informative ones for cross-modal alignment.

Image and Text Encoders

Let F^v and F^t denote the visual feature extractor and the textual feature extractor, respectively. I model F^v as a fixed object detection model (*e.g.*, Faster-RCNN), followed by a trainable fully-connected layer for mapping raw visual features in Faster-RCNN into a H-dimension joint embedding space. On the other hand, F^t encodes the word tokens in a sentence with a word embedding matrix, followed by a trainable long short-term memory (LSTM) network to model the sequential text inputs. Note that the encoders F^v and F^t are shared among $\mathcal{D}^l$ and $\mathcal{D}^u$.

The visual feature of an image v is encoded as $\boldsymbol{V} = F^v(v) = [\boldsymbol{v}_1, \cdots, \boldsymbol{v}_N] \in \mathbb{R}^{H \times N}$, where N is the maximum number of region-of-interest. Similarly, a sentence $t = [t_1, \cdots, t_M]$ is encoded as $\boldsymbol{T} = F^t(t) = [\boldsymbol{t}_1, \cdots, \boldsymbol{t}_M] \in \mathbb{R}^{H \times M}$, where M is the maximum sentence length. $(\boldsymbol{V}_i, \boldsymbol{T}_i)$ represents an annotated feature pair.

For $v^u \in \mathcal{D}^u$, I utilize an object detector (Faster RCNN (Ren et al., 2015)) to extract sequences

of regional semantics (as text tokens, $s = [s_1, \cdots, s_M]$) and generate image-semantic pairs $(\boldsymbol{V}_i^u, \boldsymbol{S}_i)$. The regional semantics are the word tokens of attribute and the class name of the objects detected from an image v^u (*e.g.*, "blue car"). The detected textual tokens are sorted by their object-wise confidence scores. I concatenate the regional semantics into one sentence, and then encode it as $\boldsymbol{S} = [\boldsymbol{s}_1, \cdots, \boldsymbol{s}_M] \in \mathbb{R}^{H \times M}$ via the shared F^t.

Given the feature representations (*i.e.*, the visual features $\boldsymbol{V}$ or the texture features $\boldsymbol{T}$), the attentive encoder can be written as (I take visual features as an example):

$$E^v(\boldsymbol{V}) = [\boldsymbol{W}_0^v \boldsymbol{V}^\top, \boldsymbol{W}_1^v \boldsymbol{V}^\top, \cdots, \boldsymbol{W}_{K-1}^v \boldsymbol{V}^\top] \tag{7.1}$$

where

$$W_{ik}^v = \frac{\exp\left(\lambda_v \alpha_{ik}^v\right)}{\sum_{i'} \exp\left(\lambda_v \alpha_{i'k}^v\right)},$$

$$\alpha_{ik}^v = \tanh(\boldsymbol{P}_k^v \frac{1}{M^v} \sum_{i'} \boldsymbol{v}_{i'})^\top \tanh(\boldsymbol{Q}_k^v \boldsymbol{v}_i).$$

The $\boldsymbol{W}_k^v \in \mathbb{R}^{1 \times M^v}$, and $\boldsymbol{P}_k^v, \boldsymbol{Q}_k^v \in \mathbb{R}^{K' \times H}$, $k \in \{0, 1, \cdots, K-1\}$ are the parameters of the attentive encoder E^v, *i.e.*, $\theta_{v-attn} = \{(\boldsymbol{W}_k^v, \boldsymbol{P}_k^v, \boldsymbol{Q}_k^v) | k \in \{0, 1, \cdots, K-1\}\}$. The λ_v is a constant temperature for the softmax function. The attentive encoder for the textural features (denoted by $E^t(\boldsymbol{T})$) works the same way but with independent parameters $\theta_{t-attn} = \{(\boldsymbol{W}_k^t, \boldsymbol{P}_k^t, \boldsymbol{Q}_k^t) | k \in \{0, 1, \cdots, K-1\}\}$. Note that E^t and E^v are shared among $\mathcal{D}_l$ and $\mathcal{D}_u$.

Thus, for an image v or v^u, the instance-level feature representation can be extracted and selectively encoded through $G^v = E^v \circ F^v$. Correspondingly, for the text description t or s, the instance-level feature can be achieved by $G^t = E^t \circ F^t$. I use $\theta_v = \{\theta_{v-attn}, \theta_{v-enc}\}$ and $\theta_t = \{\theta_{t-attn}, \theta_{t-enc}\}$ to denote the trainable parameters of G^v and G^t, respectively.

Triplet Alignment

For learning the joint embedding space, I apply a hinge-based triplet ranking loss with hard negative mining as in (Faghri et al., 2018) to align instance-wise paired visual-textual representations. Let $(\boldsymbol{a}, \boldsymbol{b})$ denotes a sampled image-text or image-semantic pair and $S(\boldsymbol{a}, \boldsymbol{b})$ is the cosine similarity. Let $\hat{\boldsymbol{b}} = \underset{\boldsymbol{b}^-}{\text{argmax}}\, S(\boldsymbol{a}, \boldsymbol{b}^-)$ and $\hat{\boldsymbol{a}} = \underset{\boldsymbol{a}^-}{\text{argmax}}\, S(\boldsymbol{a}^-, \boldsymbol{b})$ denote the hard negatives in the sampled batch. The triplet objective can be written as:

$$\begin{aligned} \ell^p(\mathcal{A}, \mathcal{B}; \alpha) = \frac{1}{L} \sum_{i=1}^{L} \{ &\left[\alpha - S(\boldsymbol{a}_i, \boldsymbol{b}_i) + S(\boldsymbol{a}_i, \hat{\boldsymbol{b}})\right]_+ \\ &+ \left[\alpha - S(\boldsymbol{a}_i, \boldsymbol{b}_i) + S(\hat{\boldsymbol{a}}, \boldsymbol{b}_i)\right]_+ \}, \end{aligned} \tag{7.2}$$

$$\ell^{NCE}(\mathcal{A}, \mathcal{B}; \theta) = -\log \frac{\exp(\text{sim}(\mathbf{a}_i, \mathbf{b}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{I}_{[k \neq i]} \exp(\text{sim}(\mathbf{a}_i, \mathbf{b}_j)/\tau)} \tag{7.3}$$

where $|\mathcal{A}| = |\mathcal{B}| = L$, $[.]_+ = max(0, .)$, and α is the margin between the similarity of positive pair and that of hard-negative pair. Since annotated image-text pairs sampled from $\mathcal{D}^l$ are more reliable than image-semantic pairs sampled from $\mathcal{D}^u$, I differentiate the strong supervision by the former from the later with a hyper-parameter β. I model the triplet alignment objective as:

$$\ell^{tri} = \beta \ell^p(G^v(v), G^v(t); \alpha_{vt}) + (1 - \beta)\ell^p(G^v(v^u), G^t(s); \alpha_{vs}) \tag{7.4}$$

Attentive Domain Alignment

A3VSE takes four different types of data, *i.e.*, $\boldsymbol{V}$, $\boldsymbol{T}$, $\boldsymbol{V}^u$, $\boldsymbol{S}$ which are regarded as samples from four different domains. As shown in Fig. 7.3, I propose using adversarial training to minimize the domain gaps among them. Specifically, I introduce six domain discriminators which are parameterized by θ_{vv^u}, θ_{ts}, θ_{vt}, $\theta_{v^u s}$, θ_{vs}, and $\theta_{v^u t}$. On the one hand, they are trained to classify samples into correct domains. On the other hand, I employ the gradient reversal layer (GRL) (Ganin and Lempitsky, 2015) to the reverse the gradients propagated from these discriminators to update G^v and G^t to minimize the domain discrepancy. Such an adversarial process can effectively diminish the discrepancy across different domains. Generally, the adversarial loss for aligning two domains is

$$\ell^d(\mathcal{A}, \mathcal{B}; \theta) = \frac{1}{|\mathcal{A}|} \sum_{i=1}^{|\mathcal{A}|} \log D_\theta(\boldsymbol{a}_i) + \frac{1}{|\mathcal{B}|} \sum_{j=1}^{|\mathcal{B}|} \log(1 - D_\theta(\boldsymbol{b}_j)) \tag{7.5}$$

where D_θ is the domain discriminator parameterized by θ. The $\mathcal{A} = \{\boldsymbol{a}_i\}$ and $\mathcal{B} = \{\boldsymbol{b}_j\}$ are the mini-batch data sampled from two domains. The instantiations of $\boldsymbol{a}$ and $\boldsymbol{b}$ can be either two of $\{G^v(v), G^v(v^u), G^t(t), G^t(s)\}$. As shown in Fig. 7.3, I perform three types of domain alignments, *i.e.*, *intra-modal alignment*, *Cross-modal alignment*, and *Transitive alignment*, which are specified in the following.

Intra-modal Alignment handles the domain gaps between the annotated and un-annotated images, and annotated text descriptions and sequences of regional semantics. Specifically,

$$\begin{aligned} \ell^{intra} &= \lambda_{vv^u} \ell^d(G^v(v), G^v(v^u); \theta_{vv^u}) \\ &+ \lambda_{ts} \ell^d(G^t(t), G^t(s); \theta_{ts}) \end{aligned} \tag{7.6}$$

Cross-modal Alignment aims at aligning the distribution of attended visual and textual features for annotated image-text pairs and inferred image-semantic pairs. That is,

$$\begin{aligned}\ell^{cross} &= \lambda_{vt}\ell^{d}(G^{v}(v), G^{t}(t); \theta_{vt}) \\ &+ \lambda_{v^{u}s}\ell^{d}(G^{v}(v^{u}), G^{t}(s); \theta_{v^{u}s})\end{aligned} \tag{7.7}$$

Transitive Alignment Transitive alignment minimizes the domain gap between annotated images and sequences of regional semantics, and the domain gap between un-annotated images and annotated text descriptions:

$$\begin{aligned}\ell^{trans} &= \lambda_{vs}\ell^{d}(G^{v}(v), G^{t}(s); \theta_{vs}) \\ &+ \lambda_{v^{u}t}\ell^{d}(G^{v}(v^{u}), G^{t}(t); \theta_{v^{u}t})\end{aligned} \tag{7.8}$$

The overall adversarial objective for the attentive alignment is:

$$\ell^{adv} = \ell^{intra} + \ell^{cross} + \ell^{trans} \tag{7.9}$$

And the final objective can be formalized as

$$\ell^{A3VSE} = \ell^{adv} + \ell^{tri} \tag{7.10}$$

Training and Inference. A min-max optimization is performed between the domain discriminators and attentive encoders:

$$\begin{aligned}(\theta_{v}, \theta_{t}) &= \text{argmin}_{\theta_{v},\theta_{t}}\ell^{A3VSE}(\boldsymbol{\theta}) \\ (\theta_{adv}) &= \text{argmax}_{\theta_{adv}}\ell^{A3VSE}(\boldsymbol{\theta}),\end{aligned} \tag{7.11}$$

where $\theta_{adv} \triangleq (\theta_{vt}, \theta_{v^{u}s}, \theta_{ts}, \theta_{vv^{u}}, \theta_{vs}, \theta_{v^{u}t})$. In each iteration, I sample a mini-batch of (v, t) from $\mathcal{D}^{l}$ and (v^{u}, s) from $\mathcal{D}^{u}$ then follow the common practice in (Ganin and Lempitsky, 2015) of adversarial training with GRL to optimize Eq. 7.10. At the inference stage, I extract the visual embedding for image v and textual embedding for sentence t through G^{v} and G^{t}.

$$\begin{aligned}(\theta_{v}, \theta_{t}) &= \text{argmin}_{\theta_{v},\theta_{t}}\ell^{NCE}(\mathcal{V}, \mathcal{V}) + \ell^{NCE}(\mathcal{T}, \mathcal{T}) + \ell^{NCE}(\mathcal{V}, \mathcal{T}) - \ell^{ADV}(\mathcal{T}, \mathcal{T}') \\ (\theta_{adv}) &= \text{argmin}_{\theta_{adv}}\ell^{ADV}(\mathcal{T}, \mathcal{T}'),\end{aligned} \tag{7.12}$$

Discussion. In A3VSE, attentive encoders and adversarial alignment cooperate to learn satisfactory visual-semantic embeddings. On the one hand, attentive encoders emphasize the informative part of the visual regions or textual entities, which helps adversarial training avoid misalignment and learn more discriminative features; on the other hand, adversarial alignment contributes to the improvement of attention mechanism of the attentive encoders in individual modalities which otherwise may be biased by the less amount of parallel image-text data.

7.2.5 Empirical Evaluation

I perform extensive experiments to confirm the superiority of the proposed A3VSE model over competitive baselines with sparsely annotated multimodal corpora. I evaluate the learned visual-semantic embeddings in cross-modal retrieval tasks on two standard benchmark datasets (Flickr30K (Young et al., 2014) and MS-COCO (Lin et al., 2014)).I constrain the amount of image-text annotations available in the training phase as an analogy to real-world scenarios where annotations are typically sparsely available.

Sparse Flickr30K. For learning with limited parallel pairs on Flickr30K (Young et al., 2014), I randomly shuffle once and trim the training set into 14,500 (50%), 5,800 (10%), and 2,900 (10%) subset of images. I sample 1, 2, and 5 text descriptions corresponding to those images. The resulting sparse training set is with size 2,900 (2%) to 72,500 (50%) out of 145,000 (100%) training image-text pairs in the original training split. The statistics of the new training splits of sparse Flickr30K can be found in Table 7.1. The standard validation and testing are used for model selection and testing.

Sparse MS-COCO. For MS-COCO (Lin et al., 2014), I follow the same procedure as performed in Flickr30K and sample 5,664 (5%), 11,382 (10%), and 22,657 (20%) images along with 1, 2, 5 corresponding text descriptions. The statistics and the number of training pairs can be found in Table 7.3. I report the testing performance on the whole 5,000 testing set.

Experimental Setup

I focus on the text-to-image retrieval task (searching images with a natural language description as the query) and the image-to-text retrieval task (searching sentences with a query image) with the learned visual-semantic embeddings. I train models under different levels of training sparsity. Model selection and testing are with the full validation and the full testing set, respectively.

For all the baselines, I use their best single model settings and the code from their publicly available Github repositories. Since there are much less paired training instances in the sparsely annotated dataset, for fair comparison and in the prevention of under-fitting, I either keep the number of (mini-batch) training iterations as 50% iterations of the full dataset or extend the training epoch by 1.2x (for 50% annotations), 2.0x (20% annotations) and 2.5x (10% annotations). Early stopping and learning rate adjustments in the baselines follow the same adjustment if feasible.

Unsupervised baseline with image-level semantics: I build an unsupervised cross-modal retrieval baseline using *NO* parallel annotations. Image-level semantics (*i.e.*,, global semantics) of

each image are extracted using pre-trained models from the following datasets:

1. Open Image (Kuznetsova et al., 2018): 5,000 semantics trained on 9 million images.
2. ImageNet Shuffle (Mettes et al., 2016), 12,073 classes defined in ImageNet.
3. Place365 (Zhou et al., 2014): 365 visual scene types.
4. Google Sports (Karpathy et al., 2014b): 478 sport-related semantics.

I remove duplicated semantic concepts, normalize the scores, and then merge them into a 16500-dimension global semantic vector $\boldsymbol{s}_g$ for each image. Each dimension can be referred to as a semantic concept in the original dataset—for example, an "aquarium" in Place365.

For retrieval, I directly match image-level semantics (tags) to text. Specifically, I expand the tokens in a sentence with the synsets defined in WordNet(Fellbaum, 1998) and construct a 16500-dimension k-hot query vector $\boldsymbol{q}$, where k is the number of matched concepts. The matching score is calculated as $r = \boldsymbol{s}_g^T \boldsymbol{q}$. This can be viewed as a simple unsupervised baseline using no parallel annotations.

Implementation Details: For regional semantics in un-annotated images, I use the Faster RCNN model in (Anderson et al., 2018) fine-tuned on Visual Genome (Krishna et al., 2017b) to extract English attribute names and class names of the objects detected from an image. Specifically, for every un-annotated image $v_j^u \in \mathcal{D}_u$, I generate $s_j = [s_{j1} \| s_{j2} \cdots \| s_{j|ROI|}]$ where "$\|$" is concatenation and $s_k = [\text{Attribute}_k \ \text{Class}_k]$ (*e.g.*, "blue car"). There are 2,000 detectable objects and attributes. These regional semantics are then sorted by the confidence scores and concatenated as a text sequence. I group the image and the sequence and encode them as an image-semantic pair $(\boldsymbol{V}^u, \boldsymbol{S})$.

In the proposed model, I set the embedding dimension H to 512. All the context vectors share the same dimension in the attention modules. Similar to (Zheng et al., 2017), I initialize word embeddings with pre-trained Glove embeddings (Pennington et al., 2014). Other hyper-parameters are set as follows: $K = 3, \alpha_{vt} = 0.2$, $\alpha_{vs} = 0.3$, $\beta = 0.8$, and $\gamma = 2/(1 + exp(-\eta p)) - 1$ as in (Ganin and Lempitsky, 2015) where $\eta = 10$ and p is linearly increased from 0 to 1 in proportional to the training epoch. The hyper-parameters for the adversarial object is set as: Intra-modal alignments:$\lambda_{vv^u} = 0.2, \lambda_{ts} = 0.1$; Cross-modal alignments:$\lambda_{vt} = 0.5, \lambda_{v^u s} = 0.5$; Transitive alignments:$\lambda_{v^u t} = \lambda_{vs} = 0.3$.

Results on Sparse Flickr30K

Table 7.1 shows the testing results with various levels of training sparsity on Flickr30K. Comparing the performance under the same percentage of annotations, the first interesting observation is

Sparse Flickr30K				Ours (A3VSE)						SCAN (Lee et al., 2018) (SOTA)					
%	#	%	# Ann	Text-to-Image			Image-to-Text			Text-to-Image			Image-to-Text		
Img	Sent	Ann	Pairs	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
10%	1/5	2%	2,900	20.7	46.0	58.5	27.6	56.2	68.1	2.0	7.2	11.7	5.1	16.0	22.9
10%	2/5	4%	5,800	28.1	55.6	66.9	42.0	69.7	79.0	16.1	35.7	46.5	18.9	39.8	53.9
10%	5/5	10%	14,500	32.0	60.1	71.0	46.8	72.8	80.7	24.6	48.1	59.3	25.9	56.3	70.9
20%	1/5	4%	5,800	29.1	56.4	68.1	43.3	71.0	81.8	17.2	37.5	47.5	21.2	44.4	55.0
20%	2/5	8%	11,600	32.6	61.6	72.3	44.8	72.7	82.8	28.4	54.0	64.6	39.0	68.0	78.6
20%	5/5	20%	29,000	34.9	64.4	73.6	48.4	77.0	85.1	29.3	56.9	68.3	42.1	71.8	81.3
50%	1/5	10%	14,500	36.7	65.1	75.9	51.6	78.7	85.7	29.5	56.3	67.3	40.2	72.2	81.4
50%	2/5	20%	29,000	42.9	70.5	80.3	61.4	83.7	89.4	33.9	61.3	71.4	46.8	75.2	84.5
50%	5/5	50%	72,500	44.5	73.8	83.3	60.9	85.7	91.6	39.2	67.5	77.2	52.6	80.3	87.5

Table 7.1: Performance comparison on the 1K testing set of Flickr30K. The models are trained with the sparsely annotated training data as specified in the left column. *% Img* stands for the percentage of training images available compared to original training images in Flickr30K. *# Sent* stands for the number of paired text descriptions available for each image. *%/# Ann* is the percentage/number of annotations used for training compared to the complete training annotations.

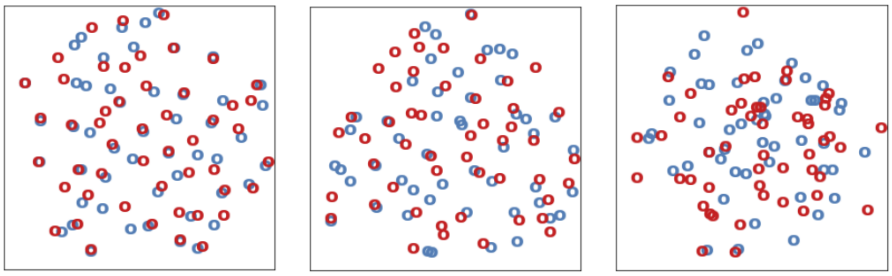
that generally speaking, it is preferred to have diverse images annotated than annotating a small number of images with more text descriptions. With the same 10% annotations, it is better to annotate 50% of images with one sentence each than 10% of images with five sentences. These results suggest that regarding data collection and annotation, visual diversity is likely to be more critical than textual diversity. Two cases of t-SNE visualization of the learned embedding are shown in Fig. 7.4a and Fig. 7.4b.

Under all sparse training set settings, the proposed model outperforms the current state-of-the-art cross-modal retrieval model (Lee et al., 2018) by a significant margin. Namely, 4.2 to 18.7 in R@1, 6.3 to 38.8 in R@5, and 5.3 to 46.8 in R@10 text-to-image retrieval tasks. Notably, a greater improvement over the current best model is achieved when less pairwise annotations are available. The improvements converge (but still outperforms) with more annotations available. A similar trend can be observed for the image-to-text retrieval task. These results demonstrate that the proposed A3VSE model can judiciously use regional semantics from un-annotated images for training its encoders and effectively learn the visual-semantic embeddings.

As shown in Table 7.2, in comparison to other recent models DAN (Nam et al., 2017), DPC (Zheng et al., 2017), and VSE++ (Faghri et al., 2018), the proposed model significantly outperforms them in all scenarios. In terms of reducing annotation effort, the proposed A3VSE model achieves competitive performance (with the criteria defined as R@10 ¿ 80.0%) trained on only 20% annotations (23,200 pairs).

	Text-to-Image			Image-to-Text		
Model	R@1	R@5	R@10	R@1	R@5	R@10
Flickr30K **0%** Ann, 0 pairs						
s_g baseline	10.5	21.5	29.2	12.1	24.0	31.1
Flickr30K 10% Img, 5/5 Sent, **10%** Ann, 14,500 pairs						
DPC (Zheng et al., 2017)	8.5	26.0	40.9	11.8	45.5	66.0
DAN (Nam et al., 2017)	10.1	25.3	42.8	12.2	41.7	64.5
VSE++ (Faghri et al., 2018)	7.2	27.5	40.5	10.5	40.2	62.8
SCAN (Lee et al., 2018)	24.6	48.1	59.3	25.9	56.3	70.1
Ours (A3VSE)	**32.0**	**60.1**	**71.0**	**46.8**	**73.2**	**80.7**
Flickr30K 50% Img, 2/5 Sent, **20%** Ann, 29,000 pairs						
DPC (Zheng et al., 2017)	26.4	53.0	63.9	35.8	68.5	79.7
DAN (Nam et al., 2017)	26.9	52.3	64.8	37.2	69.9	78.2
VSE++ (Faghri et al., 2018)	27.3	54.5	66.0	33.5	65.2	78.2
SCAN (Lee et al., 2018)	33.9	61.3	71.4	46.8	75.2	84.5
Ours (A3VSE)	**42.9**	**70.5**	**80.3**	**61.4**	**83.7**	**89.4**
Flickr30K **100%** Ann, 145,000 pairs						
DPC (Zheng et al., 2017)	39.1	69.2	80.9	55.6	81.9	89.0
DAN (Nam et al., 2017)	39.4	69.2	79.1	55.0	81.8	89.5
VSE++ (Faghri et al., 2018)	39.6	70.1	79.8	53.1	82.1	87.5
SCAN (Lee et al., 2018)	45.8	74.4	83.0	61.8	87.5	93.7
Ours (A3VSE)	**49.5**	**79.5**	**86.6**	**65.0**	**89.2**	**94.5**

Table 7.2: Performance comparison with baselines on two sparse settings in Flickr30K.



(a) 20% Img, 5/5 Sent (b) 10% Img, 5/5 Sent (c) No s in (b)

Figure 7.4: t-SNE visualization of the embedded testing images (blue) and sentences (red) under sparse Flick30K. Paired ones are expected to be close to each other.

Sparse MS-COCO				Ours (A3VSE)						SCAN (Lee et al., 2018) (SOTA)					
%	#	%	# Ann	Text-to-Image			Image-to-Text			Text-to-Image			Image-to-Text		
Img	Sent	Ann	Pairs	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
5%	1/5	1%	5,664	14.2	35.8	48.9	19.2	44.2	57.4	9.0	24.4	35.0	9.5	27.2	38.9
5%	2/5	2%	11,328	16.1	39.5	52.8	22.2	47.8	61.8	12.7	31.6	42.9	11.9	33.1	46.1
5%	5/5	5%	28,320	19.7	44.4	57.7	27.8	55.9	68.8	16.8	40.0	52.6	21.0	47.3	61.2
10%	1/5	2%	11,328	17.7	41.9	54.8	24.6	51.5	63.7	12.7	31.8	43.2	12.8	34.1	48.2
10%	2/5	4%	22,656	20.3	45.5	58.8	26.5	55.6	68.8	17.3	41.5	54.4	22.4	49.7	62.5
10%	5/5	10%	56,640	23.2	50.5	64.1	30.5	60.4	73.1	19.4	44.3	57.3	25.5	53.8	67.6
20%	1/5	4%	22,657	20.0	45.9	59.5	26.9	54.4	67.9	16.3	37.9	50.3	17.8	43.4	57.0
20%	2/5	8%	45,314	24.5	51.8	64.8	32.4	63.0	75.1	20.3	44.5	57.3	24.2	53.7	67.5
20%	5/5	20%	113,287	27.4	56.0	68.9	38.3	68.1	79.3	21.1	45.2	57.8	24.2	54.8	68.6

Table 7.3: Performance comparison on the 5K testing set of MS-COCO.

It is noteworthy that the unsupervised approach with global semantics which uses *NO* image-text pairs cannot deliver satisfactory retrieval performance when queried with natural language, indicating that there is a clear domain shift between the semantic pool of current image classification/ tagging models and the natural language queries. A similar phenomenon is observed in the ablation studies. Moreover, from the crossover of 10.5 R@1 in Fig. 7.2, the unsupervised global semantics from external classification datasets is worth as many as 14,000 image-text annotation pairs for the recent cross-modal retrieval models. Notably, A3VSE achieves 29.1 R@1 even trained with only 5,800 pairs.

Results on Sparse MS-COCO

Table 7.3 shows the results on the harder 5K testing set of MS-COCO. I sample 5%, 10%, 20% of images in MS-COCO to keep the number of training pairs more comparable to Flickr30K. The proposed model delivers the best performance on most metrics under all sparsity settings. For text-to-image retrieval, it outperforms SCAN (Lee et al., 2018) by 2.9 to 6.3 in R@1, 4.0 to 11.4 in R@5, and 4.4 to 13.9 in R@10. A similar trend can be observed in image-to-text retrieval task. The comparison with other recently published models is shown in Table 7.4, where the proposed model achieves the best performance in all sparse corpus scenarios.

Despite using only 20% of image-text annotations, the proposed model still achieves competitive performance (with the criteria defined as R@10 ¿ 70.0%) in the more challenging 5K testing set in MS-COCO. More than 80% of annotation effort for the image-text pairs could potentially be relieved. Based on the quantitative results on multiple datasets, I validate the superiority and the annotation efficiency of the proposed A3VSE model.

	Text-to-Image			Image-to-Text		
Model	R@1	R@5	R@10	R@1	R@5	R@10
MS-COCO **0%** Ann, 0 pairs						
s_g baseline	7.5	16.8	23.2	8.8	15.0	22.8
MS-COCO 10% Img, 1/5 Sent, **2%** Ann, 11,328 pairs						
DPC (Zheng et al., 2017)	8.1	28.3	38.0	10.5	30.8	41.0
DAN (Nam et al., 2017)	8.8	28.3	37.1	11.1	30.1	42.5
VSE++ (Faghri et al., 2018)	8.5	27.6	36.5	10.7	30.2	44.5
SCAN (Lee et al., 2018)	12.7	31.8	43.2	12.8	34.1	48.2
Ours (A3VSE)	**17.7**	**41.9**	**54.8**	**24.6**	**51.5**	**63.7**
MS-COCO 50% Img, 2/5 Sent, **20%** Ann, 113,287 pairs						
DPC (Zheng et al., 2017)	19.1	41.0	55.5	20.5	45.1	60.2
DAN (Nam et al., 2017)	19.5	40.8	54.0	20.7	47.7	61.7
VSE++ (Faghri et al., 2018)	19.5	41.2	56.5	21.5	48.5	63.5
SCAN (Lee et al., 2018)	22.3	47.5	60.2	25.5	56.1	70.5
Ours (A3VSE)	**28.2**	**57.9**	**70.6**	**38.4**	**69.5**	**81.1**
MS-COCO **100%** Ann, 566,435 pairs						
DPC (Zheng et al., 2017)	25.3	53.4	66.4	41.2	70.5	81.1
DAN (Nam et al., 2017)	29.8	58.8	70.0	40.8	70.0	79.8
VSE++ (Faghri et al., 2018)	30.3	56.0	72.4	41.3	69.5	81.2
SCAN (Lee et al., 2018)	34.4	63.7	75.7	46.4	77.4	87.2
Ours (A3VSE)	**39.0**	**68.0**	**80.1**	**49.3**	**81.1**	**90.2**

Table 7.4: Performance comparison with baselines on two sparse settings in MS-COCO.

Ablation Study

To quantify the contribution from individual components, I conduct ablation studies evaluating the cross-modal retrieval performance with models trained with 10% of images and 5/5 corresponding text descriptions (10% annotations) in Flickr30K. In each experiment, I remove one or change a component of concern to quantify its relative importance. A component is more important with a more substantial drop. For the experiment without semantics ($\boldsymbol{s}$), I remove all the regional semantics from the input and show the performance of the vanilla model. Then I swap the sequence of regional semantics with global semantics $\boldsymbol{s}_g$ and encode global semantics (can be viewed as image-level tags after applying a 0.3 threshold) with the shared word embedding matrix.

For the internal modules and adversarial objectives, I either remove the attention layer with mean pooling over encoded visual/textual entities as the final instance-level representation, or I purge an adversarial objective from Eq. 7.10 during the training phase.

Flickr30K 10% Img 5/5 Sent, **10**% Ann, 14,500 pairs						
	Text-to-Image			Image-to-Text		
Model	R@1	R@5	R@10	R@1	R@5	R@10
No $\boldsymbol{s}$	23.4	47.9	58.2	26.5	58.1	71.5
Swap $\boldsymbol{s}$ with $\boldsymbol{s}_g$	29.0	56.3	67.2	40.5	67.4	77.6
$\boldsymbol{s}$, without attention	23.8	50.1	62.7	35.8	64.3	75.1
$\boldsymbol{s}$, without L_{adv}	30.9	58.5	69.0	43.8	70.9	79.5
Without ℓ^{intra}	31.8	59.6	**71.0**	44.8	72.5	**80.8**
Without ℓ^{cross}	31.3	59.2	70.5	45.2	71.8	80.1
Without ℓ^{trans}	31.5	59.7	70.9	46.1	71.8	80.3
Full model	**32.0**	**60.1**	**71.0**	**46.8**	**72.8**	80.7

Table 7.5: Ablation study of the proposed model

Table 7.5 shows the results of the ablation study. I observe that while global semantics boost model performance from the vanilla model, the regional semantics is the better choice even if they have a relatively small vocabulary size (1,104 versus 1,576) for the un-annotated images in sparse Flickr30K. The visualization of learned embeddings in Fig. 7.4b and Fig. 7.4c double confirms the difference. One possible explanation for this phenomenon is that regional semantics are more similar to natural language descriptions. I observe that the distribution of vocabulary is closer (13.1% Intersection over Union (IoU)) between the natural language queries and the regional semantics than the global semantics (9.8% IoU). For instance, in natural language descriptions, people tend to describe an image with "frog" or "dog" rather than the detected global semantics "Amphibian" and "Havanese".

Additionally, the attentive adversarial learning with domain discriminators plays a vital role in closing the domain gaps between annotated and un-annotated inputs, delivers improved performance over models without adversarial objectives. However, I observe small variants among the best metrics over various configurations, suggesting that careful hyper-parameter tuning may be required to achieve the optimal performance. I leave the robust automatic tuning for aligning multiple heterogeneous domains as the future work.

(a) 50% Img, 2/5 Sent, 29,000 pairs (b) 10% Img, 2/5 Sent, 5,800 pairs (c) Failures of (b)

Figure 7.5: Qualitative examples of the proposed A3VSE model in text-to-image retrieval task (the upper two rows) and image-to-text retrieval task (the bottom row) on Flickr30K.

Qualitative Results

Fig. 7.5 illustrates sampled qualitative testing results in the image-to-text and text-to-image retrieval tasks on sparse Flickr30K. The top two rows show the top four retrieved images given the natural language query above. The one and only one correct image is marked in green or red if rank is greater than 10. The image-to-text retrieval results are depicted in the bottom row. I list the top five retrieved sentences and the corresponding query image. The correct sentences (up to five) are colored in green, otherwise red.

In most cases, the proposed model generates satisfactory results. As less parallel image-text pairs are available for training, I observe performance degeneration. For the failure cases, as expected, I observe that many failures result from out-of-vocabulary words (*e.g.*, "amplifier" and "harp") in the sentences.

Summary

To reduce expensive human annotation cost, I have presented a novel annotation efficient A3VSE model for learning improved visual-semantic embeddings (VSEs) with sparsely annotated multi-modal corpora. The proposed model jointly leverages strong supervision from image-text pairs and weak supervision from image-semantic pairs where the regional semantics are extracted from the un-annotated image collection. To further unify the heterogeneous inputs in the joint embedding space, the proposed model employs attention-enhanced adversarial objectives to model intra-modal, cross-modal, and transitive alignment to selectively align annotated and the un-annotated portion of visual and textual inputs.

In sparse Flickr30K and MS-COCO, the proposed model consistently and significantly outperforms recent competitive baselines. In comparison to global semantic tags, I have shown that regional semantics are more feasible for learning VSEs under sparsity. Concerning reducing annotation effort, I have presented insights towards efficient annotation collection and utilization. I have demonstrated that nearly 80% of the annotations can be reduced with the proposed model while achieving competitive results to recent models trained with the complete annotations.

7.3 Multimodal Machine Translation

7.3.1 Motivation

Most of the machine translation tasks only focus on textual sentences of the source language and target language; however, in the real world, the sentences may contain information about what people see. Beyond the bilingual translation, with additional information from images, I would further resolve the problem of ambiguity in languages. For example, the word "*bank*" may refer to the financial institution or the land of the river's edge. It would be confusing if I only look at the language itself. In this task, the image may help to disambiguate the meaning if it shows that there is a river, and thus the "*bank*" means "*river bank*".

In this work, as one of the earliest work in multimodal machine translation (Elliott et al., 2016; Huang et al., 2016; Elliott and Kádár, 2017), I explore three approaches to integrating multimodal information (text and image) into the attention-based encoder-decoder architecture for learning a joint multilingual-VSE space. I transform and make the visual features as one of the steps in the encoder as text, and then make it possible to attend to both the text and the image while decoding. The image features I used are (visual) semantic features extracted from the entire images (global) as well as the regional bounding boxes proposed by the region-based convolutional neural networks (Faster RCNN) (Ren et al., 2015). In our empirical evaluation, I verify the benefit of visual information over text-only machine translation. I also observe the superiority of utilizing regional semantics from visual objects with proper architecture modifications on the vanilla encoder-decoder framework.

7.3.2 Prior Work

In fields of machine translation, neural networks attract lots of research attention recently that the encoder-decoder framework is widely used Sutskever et al. (2014b). Nevertheless, the main

drawback of this neural machine translation (NMT) framework is that the decoder only depends on the last state of the encoder, which may deteriorate the performance when the sentence is long. As a solution, the attention-based encoder-decoder framework as shown in Figure 7.6 is proposed Bahdanau et al. (2015); Luong et al. (2015a). With the attention model, in each time step, the decoder depends on both the previous LSTM hidden state and the context vector, which is the weighted sum of the hidden states in the encoder[1].

Multimodal machine translation (MMT) is firstly introduced in (Specia et al., 2016) as a multi-encoder single-decoder framework with additional image inputs. MMT aim to utilize additional visual information as the complimentary information source to improve translation quality. My work Huang et al. (2016) is one of the first research efforts on this task at that time. Instead of relying on whole-image visual features, I seek to leverage object-level visual information and study method incorporating multimodal attention in to text-only machine translation.

7.3.3 Multimodal Attention in Machine Translation

In this work, I investigate three methods to incorporate the regional visual feature into the encoder-decoder framework for multimodal neural machine translation. Based on the encoder-decoder framework, the attention-based model aims to handle the missing order and source information problems in the basic encoder-decoder framework. At each time step t in the decoding phase, the attention-based model attends to subsets of words in the source sentences that can form up the context, which can help the decoder to predict the next word. This model infers a variable-length alignment weight vector $\mathbf{a_t}$ based on the current target state $\mathbf{h}_t$ and all source states $\mathbf{h}_s$. The context feature vector $\mathbf{c}_t = \mathbf{a}_t \cdot \mathbf{h}_s$ is the weighted sum of the source states $\mathbf{h}_s$ according to $\mathbf{a_t}$, which is defined as:

$$\mathbf{a}_t(s) = \frac{e^{score(\mathbf{h}_t, \mathbf{h}_s)}}{\sum_s' e^{score(\mathbf{h}_t, \mathbf{h}_s')}} \tag{7.13}$$

The scoring function $score(\mathbf{h}_t, \mathbf{h}_s)$ can be referred to as a content-based measurement of the similarity between the currently translating target and the source words. I utilize a transformation matrix $\mathbf{W}_a$ which associates source and target hidden state of learning the general similarity measure by:

$$score(\mathbf{h}_t, \mathbf{h}_s) = \mathbf{h}_t \mathbf{W}_a \mathbf{h}_s \tag{7.14}$$

I produce an attentional hidden state $\hat{\mathbf{h}}_t$ by learning $\mathbf{W}_c$ of a single layer perceptron activated by $tanh$. The input is simply the concatenation of the target hidden state $\mathbf{h}_t$ and the source-side

[1]Readers may also check Chapter §3 for preliminary regarding attention-based encoder-decoder models in machine translation.

context vector $\mathbf{c}_t$:

$$\hat{\mathbf{h}}_t = tanh(\mathbf{W}_c[\mathbf{c}_t; \mathbf{h}_t]) \tag{7.15}$$

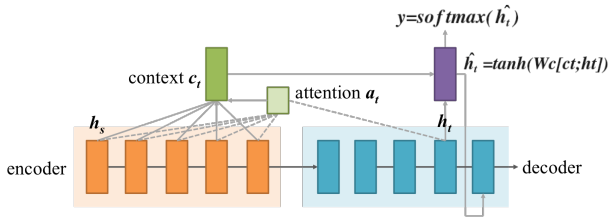


Figure 7.6: Attention-based neural machine translation framework using a context vector to focus on a subset of the encoding hidden states.

After generating the context feature vector and the attentional hidden state, the target word is predicted through the softmax layer with the attentional hidden state $\mathbf{h}_t$ vector by $p(y_t|\mathbf{x}) = softmax(\mathbf{W}_s\hat{\mathbf{h}}_t)$. In the following, I introduce how to incorporate images features based on the attention models.

Model 1: LSTM with global visual feature Visual features from the convolution neural network (CNN) may provide additional information to textual features in machine translation with multiple modalities. As depicted in Figure 7.7, I propose to append visual features at the head/tail to the original text sequence in the encoding phase. Note that for simplicity, I omit the attention part in the following figures.

Global (i.e., whole image) visual features are extracted from the last fully connected layer known as $fc7$, a 4096-dimensional semantic layer in the 19-layered VGG (Simonyan and Zisserman, 2014). With the dimension mismatch and the inherent difference in content between the visual and textual embedding, a transformation matrix $\mathbf{W}_{img}$ is proposed to learn the mapping. The encoder then encodes both textual and visual feature sequences to generate the representation for decoding. In the decoding phase, the attention model weights all the possible hidden states in the encoding phase and produce the context vector $\mathbf{c}_t$ with Eq. 7.13 and Eq. 7.14 for NMT decoding.

Model 2: LSTM with multiple regional visual features In addition to adding only one global visual feature, I extend the original NMT model by incorporating multiple regional features in the hope that those regional visual attributes would assist LSTM to generate better and more accurate representations. The illustration of the proposed model is depicted in 7.8. I will first explain how to determine multiple regions from one image and explain how these visual features are extracted and sorted.

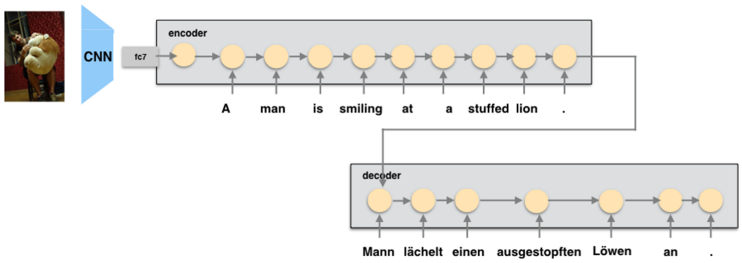


Figure 7.7: Model 1: Attention-based NMT with single additional global visual feature. Decoder may attend to both text and image steps of encoding. For clarity, the possible attention path is hidden here.

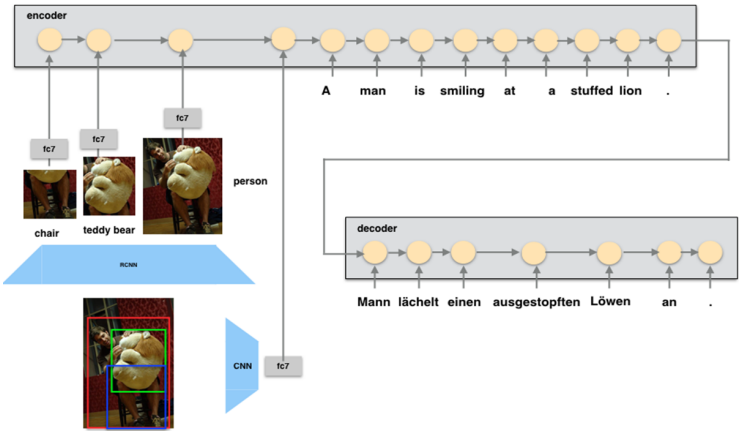


Figure 7.8: Model 2: Attention-based NMT with multiple additional regional visual features.

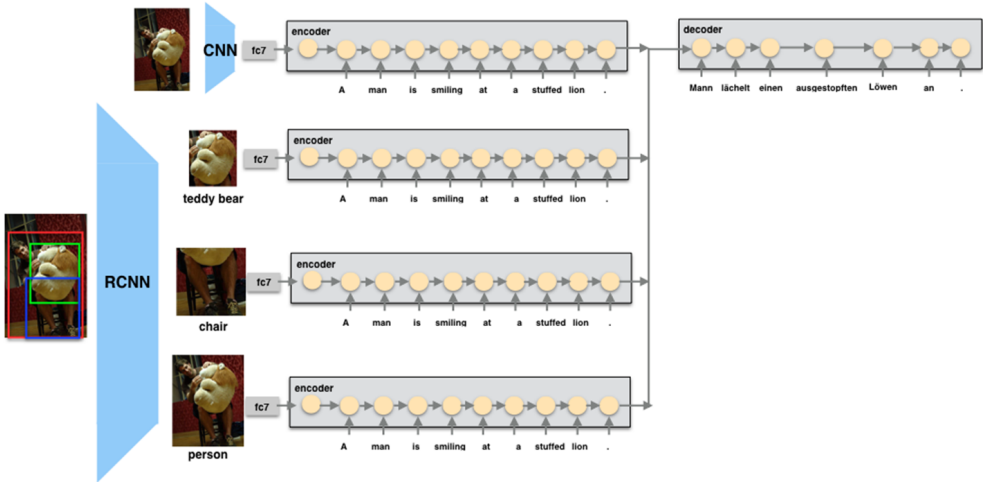


Figure 7.9: Model 3: Parallel LSTM threads with multiple additional regional visual features.

Intuitively, objects in an image are most likely to appear in both source and target sentences.

Therefore, I utilize the region proposal network (RPN) in the region-based convolutional neural network (Ren et al., 2015) (R-CNN) to identify objects and their bounding boxes in an image and then extract visual feature from those regions. In order to integrate these images into the original sequence in the LSTM model, I design a heuristic approach to sort those visual features. The regional features are fed in the ascending order of the size of the bounding boxes, followed by the original global visual feature and the text sequence. Visual features are sequentially fed in such order since important features are designed to be closer to the encoded representation. Heuristically, larger objects may be more noticeable and essential in an image described by both the source and target language contexts.

In the implementation, I choose the top four regional objects plus the whole image and then extracted their $fc7$ with VGG-19 to form the visual sequence followed by the text sequence. If there are less than four objects recognized in the original image, zero vectors are padded instead of the batch process during training.

Model 3: Parallel LSTM threads To further alleviate the assumption that regional objects share some pre-defined order, I further propose a parallel structure, as shown in Figure 7.9. The encoder of NMT is composed of multiple encoding threads where all the LSTM parameters are shared. In each thread, a (regional) visual feature is followed by the text sequence. This parallel structure would associate the text to the most relevant objects in the encoding phase and distinguish them when computing attention during decoding. Intuitively, the text sequence follows a regional object that would be interpreted as encoding the visual information with the textual description (i.e., encoding captions as well as visual features for that object). An encoder hidden state for attention can be interpreted as the "word" imprinted by the semantics features of some regional object. The decoder can, therefore, distinctively attend to words that describe different visual objects in multiple threads.

In the encoding phase, parameters in LSTM are shared over threads. All possible hidden states over multiple threads are recorded for attention. At the end of the encoding phase, the outputs of different encoding threads are fused to generate the final embedding of the whole sentence as well as all the image objects. In the decoding phase, candidates of global attention are all the text hidden states over multiple threads. For example, at time t, the decoder may choose to attend to 'bear' at the second thread (which sees a teddy bear image at the beginning) as well as the 'bear' in the global image thread. At time $t + 1$, the decoder may switch to another thread and focus on "the man" with the person image.

7.3.4 Empirical Evaluation

Experimental Setup

I follow the standard split in Multi30K (Elliott et al., 2016). Global visual features (*fc7*) are extracted with VGG-19 (Simonyan and Zisserman, 2014). For regional visual features, the region proposal network in Faster RCNN (Ren et al., 2015) first recognizes bounding boxes of objects in an image and then I computed 4096-dimensional *fc7* features from these regions with VGG-19. The RPN of Faster RCNN is pre-trained on ImageNet dataset [2] and then fine-tuned on MSCOCO dataset [3] with 80 object classes.

I use a single-layered LSTM with 256 cells and 128 batch size for training. The dimension of word embedding is 256. $\mathbf{W}_{img}$ is a 4096×256 matrix transforming visual features into the same embedding space as words. When training NMT, I follow (Luong et al., 2015b) with similar settings: (a) I uniformly initialized all parameters between -0.1 and 0.1, (b) I trained the LSTM for 20 epochs using simple SGD, (c) the learning rate was initialized as 1.0, multiplied by 0.7 after 12 epochs, (d) dropout rate was 0.8. Note that the same dropout mask and NMT parameters are shared by all LSTM threads in model 3.

	BLEU	METEOR
Text baseline	34.5 (0.7)	51.8 (0.7)
m1:image at tail	34.8 (0.6)	51.6 (0.7)
m1:image at head	35.1 (0.8)	52.2 (0.7)
m2:5 sequential RCNNs	36.2 (0.8)	53.4 (0.6)
m3:5 parallel RCNNs	**36.5** (0.8)	**54.1** (0.7)

Table 7.6: BLEU and METEOR of the proposed multimodal NMT

7.3.5 Discussion

The quantitative performance of the proposed models can be seen in Table 7.6. I evaluate BLEU and METEOR scores with tokenization under the official settings of the WMT 2016 multimodal machine translation challenge. The text-only baseline is the NMT implementation with global attention. Adding a single global visual feature from an image at the head of a text sequence improves BLEU by 0.6% and METEOR by 0.4%, respectively.

[2] http://image-net.org/
[3] http://mscoco.org/

The results show that the additional visual information improves the translations in this dataset. However, lukewarm improvement is not as significant as I expected. One possible explanation is that the information required for the multimodal translation task is mostly self-contained in the source text transcript. Adding global features from whole images do not provide extra supplementary information and thus results in a subtle improvement.

Detailed regional visual features provide extra attributes and information that may help the NMT translates better than the text-only baselines. In our experiment, the proposed model2 with multiple regional and one global visual features showed an improvement of 1.7% in BLEU and 1.6% in METEOR, while model3 showed an improvement of 2.0% in BLEU and 2.3% in METEOR. The results correspond to our observation that most sentences would describe important objects which could be identified by R-CNN. The most commonly mentioned object is "person". It is likely that the additional attributes provided by the visual features about the person in an image help to encode a more detailed context and thus benefit NMT decoding. Other high frequency objects are "car", "baseball", "cellphone", etc.

For the proposed LSTM with multiple regional visual features (model 2), the semantic features in $fc7$ of the regions-of-interest in an image provide additional regional visual information to form a better sentence representation. I also experimented with other sorting methods, including descending size, random, and categorical order, to generate the visual sequences. However, ascending-ordered sequences achieve the best result.

For the proposed parallel LSTM architecture with regional visual features (model 3), the regional visual features further help the NMT decoder to attend more accurately and accordingly to focus on the right thread where the local visual attributes twiddle the hidden states. The best result of our models achieve 36.5% in BLEU, and 54.1% in METEOR, which is comparable to the state-of-the-art Moses results in this challenge.

7.4 Unsupervised Multimodal Machine Translation via Pseudo Visual Pivoting

7.4.1 Motivation

Neural machine translation (MT) (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014a) has achieved near human-level performance (Wu et al., 2016). However, its effectiveness strongly relies on the availability of large-scale parallel corpora. Unfortunately, preparing the parallel data

remains a challenge as there are more than 6,500 languages in the world, and recruiting translators with bilingual or multilingual knowledge to cover all those languages is impractical.

As a result, developing methods alleviating the need of well-annotated large parallel corpora has recently attracted increasing attention in the community. These methods fall into two broad categories. The first type of methods use a third language as the pivot (Firat et al., 2016; Chen et al., 2017; Cheng et al., 2017; Johnson et al., 2017) to enable zero-resource translation. Although the progress is encouraging, pivoting with a third language still demands bilingual knowledge for collecting large-scale parallel source-pivot and pivot-target corpora. The second type of methods explore unsupervised approaches (Conneau et al., 2018a; Artetxe et al., 2018; Lample et al., 2018a) have recently achieved impressive translation quality. These methods rely only on monolingual data and back-translation (Sennrich et al., 2016a). However, as discussed in (Lample et al., 2018b), the alignment of source-target sentences is uncertain and highly subject to proper initialization.

Using visual content for unsupervised MT (Chen et al., 2018b; Su et al., 2019) is a promising solution for pivoting and alignment based on its availability and feasibility. Abundant multimodal content in various languages are available online (*e.g.*, Instagram and YouTube). It is also easier to recruit monolingual annotators to describe an image than to find multilingual translators to translate sentences. Importantly, visual content is eligible to improve the alignments in the language latent spaces since the physical visual perception is similar among people speaking different languages (*e.g.*, similar "blue car" for a German and a French).

Motivated by these insights, I propose a novel unsupervised multimodal MT framework incorporating images as pseudo pivots promoting latent space alignment. In addition to use features of visual objects for multimodal back-translation, I align a shared multilingual visual-semantic embedding (VSE) space via leveraging disjoint image-sentence pairs in different languages. As illustrated in Fig. 7.11, for sentences approximately pivoted by similar images (src-*img*-tgt), drawing embeddings of corresponding image-sentence pairs closer results in better alignments of semantically equivalent sentences in the language latent spaces. Inspired by back-translation, I further explore another pseudo pivoting strategy, which approximates multilingual sentence pairs (*src*-img-*tgt*) conditioned on a real image via captioning. Instead of using annotation of images for pivoting as in (Chen et al., 2018b), I generate sentences in two languages pivoted on the real image, and then approximately pairing them as weak supervision for training unsupervised MT system. This approach is analogous to a cross-modal version of back-translation.

In short, my work in this thesis makes the following contributions: (1) Building a unified view of employing visual content for pseudo pivoting. (2) Improving the alignments in the shared

multilingual multimodal embedding space for unsupervised MMT with disjoint image-text pairs in different languages. (3) The proposed model achieves state of the art on Multi30K and generalizes well to the text-only scenario.

In the following sections, I first review prior works in unsupervised multimodal MT and identify the unique contribution of the proposed *pseudo visual pivoting* method. Then I provide details of my model which employs multimodal back-translation and features pseudo visual pivoting for learning a shared multilingual visual-semantic embedding space and incorporating visually-pivoted captioning as additional weak supervision. Finally, I present the experimental results on the widely used Multi30K dataset. I show that the proposed model significantly improves over the state-of-the-art methods and generalizes well when images are not available at the testing time.

7.4.2 Prior Work

Multimodal Machine Translation. Supervised multimodal machine translation (MMT) is firstly introduced in (Specia et al., 2016) as a multi-encoder single-decoder framework with additional image inputs with an insight that the additional visual information would provide complimentary information to improve translation quality. Huang et al. (2016) encode word sequences with regional visual objects while Calixto and Liu (2017) leverage global visual feature. While these methods achieve improvements, their advantage over the text-only models is still minor under the supervised scenario. As analyzed in (Caglayan et al., 2019), visual content is more critical when the textual content is limited or uncertain in MMT.

In contrast to the view in conventional supervised MMT which considers visual information as a complimentary information source, in this thesis I envision that visual information implicitly better serve as the bridge or "pseudo" pivot for associating the source and target languages when the source-target translation pairs are limited or even absent (unsupervised MT).

Unsupervised Machine Translation. In conventional unsupervised MT, one of the proven methods is to use the third pivoting language. Chen et al. (2017) use a teacher-student framework and assume parallel sentences share a similar likelihood for generating sentences in the third language while (Cheng et al., 2017) maximize the expected likelihood. In contrast, the proposed *pseudo visual pivoting* does not rely on a third language. My work is along the line of research in Lample et al. (2018a,b); Lample and Alexis (2019), which aims at learning an aligned latent space between the two languages to translate by reconstruction. Nevertheless, I focus on the multimodal setup where the visual space is dissimilar to the language spaces with challenging asymmetric

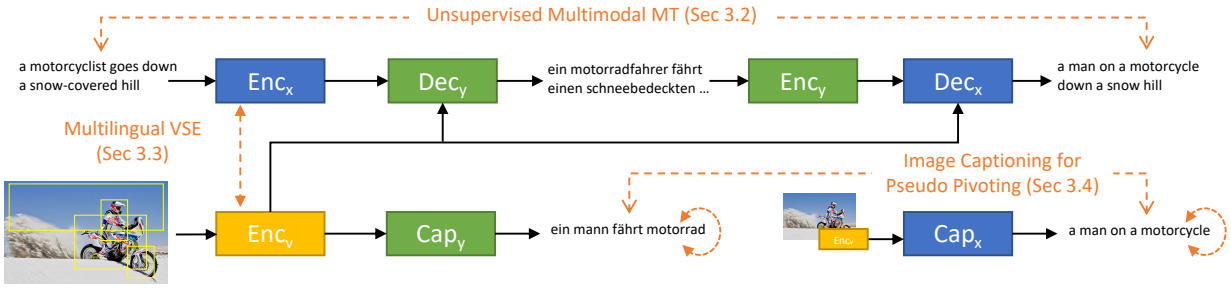


Figure 7.10: The proposed model structure (English↔German). I incorporate visual objects for unsupervised multimodal MT and improve the latent space alignments via pseudo visual pivoting with designed objectives.

interactions between modalities.

There are three recent unsupervised MMT works. Nakayama and Nishida (2017) learn modal-agnostic fixed length image/sentence embeddings. In contrast, my model promotes fine-grained (object-token) varying-length embedding, which better aligns VSE space. Game-MMT (Chen et al., 2018b) use a captioning and a translation model maximizing the likelihood of translated captions to original sentences. My model synthesize captions for symmetric back-translation and considers no ground truth image annotation in the loop. Empirically, it is preferred to separate real and generated captions. UMMT (Su et al., 2019) uses Transformers, autoencoder loss, and multimodal back-translation. My model does not rely on autoencoders. Finally, my model leverages object detection for multimodal back-translation and equips *pseudo visual pivoting*, a novel and generalized method that covers all possible scenario that use the visual space as the implicit pivoting space for aligning language sub-spaces.

7.4.3 Unsupervised Multimodal Machine Translation

As illustrated in Fig. 7.10, my model is composed of seven modules: Two encoder-decoder pairs for translation, two decoders for captioning, and one shared visual encoder. In the following, I first define the problem and detail the basic MMT model architecture under the unsupervised setup. Then I introduce my approaches for pseudo visual pivoting with multilingual VSE and pivoted captioning.

Multimodal MT

Multimodal machine translation (Specia et al., 2016) (MMT) considers additional images as a complementary information source for MT. An image $\mathbf{z}$ and the description in two languages form

a triplet $(\mathbf{x}, \mathbf{y}, \mathbf{z}) \in (\mathcal{X}, \mathcal{Y}, \mathcal{Z})$. The Transformer encoder reads the source sentence and encodes it with hierarchical self-attention into $\mathbf{h}^x = \{\mathbf{h}_1^x, \cdots, \mathbf{h}_N^x\}, \mathbf{h}_i^x \in \mathbb{R}^d$, where d is the dimension of the embedding space. The visual encoder encodes the image into $\mathbf{h}^z = \{\mathbf{h}_1^z, \cdots, \mathbf{h}_K^z\}, \mathbf{h}_i^z \in \mathbb{R}^d$, $K_{\text{max}} = 36$. Most previous work (Chen et al., 2018b; Su et al., 2019) use 2D ($K = 14 \times 14$) feature maps of ImageNet pre-trained ResNet (He et al., 2016b). In contrast, I utilize the regional features of K salient visual objects in an image extracted by Faster-RCNN (Ren et al., 2015) and a 1-layer MLP as the encoder to encode visual objects.

Various attention strategies for sequence-to-sequence learning have been addressed in (Libovický and Helcl, 2017). My model employs the hierarchical multi-head multimodal attention for decoding. For decoding at time stamp i, the textual attention $\text{Attn}(\mathbf{h}_i^y, \mathbf{h}^x)$ computes the context vector $\mathbf{c}_i = \sum_j \alpha_j \mathbf{h}_j^x$ via a attention-based alignment $\alpha_j = \text{Align}(\mathbf{h}_i^y, \mathbf{h}_j^x)$, where $\sum_j \alpha_j = 1$ and $\mathbf{h}_i^y$ is the decoder state. Essentially, the one-head attention in Transformer is implemented as $\mathbf{c}_i = \text{softmax}(\mathbf{Q}_i(\mathbf{K}^x)^\top/\sqrt{d})\mathbf{V}^x$ where $\{\mathbf{Q}, \mathbf{K}^x, \mathbf{V}^x\}$ are the packed d-dimensional *Query, Key, Value* vectors, which are the mapped and packed version of $\{\mathbf{h}_i^y, \mathbf{h}^x, \mathbf{h}^x\}$. For decoding with encoded visual and textual inputs, I utilize multimodal attention to compute the context vector $\mathbf{c}_i$:

$$\mathbf{c}_i^x = \text{Attn}(\mathbf{h}_{i-1}^y, \mathbf{h}^x) + \lambda_v \text{Attn}(\mathbf{h}_{i-1}^y, \mathbf{h}^z) \tag{7.16}$$

In practice I set $\lambda_v = 1.0$. The multimodal decoder models the likelihood to predict the next token as:

$$p(y_i|\mathbf{y}_{<i}, \mathbf{x}, \mathbf{z}) = \text{softmax}(f(\mathbf{c}_i, y_{i-1}, \mathbf{h}_{i-1}^y), \tag{7.17}$$

where $f(.)$ denotes the aggregated non-linear feature mapping in Transformer.

Unsupervised Learning

Unsupervised multimodal MT (Nakayama and Nishida, 2017; Chen et al., 2018b; Su et al., 2019) poses a new yet challenging problem. On both the source and target sides, only non-overlapping monolingual multimodal data are presented for training and validation. Specifically, the data available are: $(\mathbf{x}, \mathbf{z}_x) \in (\mathcal{X}, \mathcal{Z})$, $(\mathbf{y}, \mathbf{z}_y) \in (\mathcal{Y}, \mathcal{Z})$, such that $\{\mathbf{x}\} \cap \{\mathbf{y}\} = \phi$, $\{\mathbf{z}_x\} \cap \{\mathbf{z}_y\} = \phi$. Note that there are no parallel translation pairs available (unsupervised), and the images are mutually exclusive for different languages.

For multimodal back-translation, the generated pseudo target sentence conditioned on the source sentence and image can be re-written as $g^*(\mathbf{x}, \mathbf{z}_x) = \text{argmax}\, p_{xz\to y}(\mathbf{y}|\mathbf{x}, \mathbf{z}_x)$, where $p_{xz\to y}(\mathbf{y}|\mathbf{x}, \mathbf{z}) = \prod_{i=1}^M p(y_i|\mathbf{y}_{<i}, \mathbf{x}, \mathbf{z})$. Similar for $p_{yz\to x}(\mathbf{x}|\mathbf{y}, \mathbf{z})$ and $h^*(\mathbf{y}, \mathbf{z}_y)$. For unsupervised

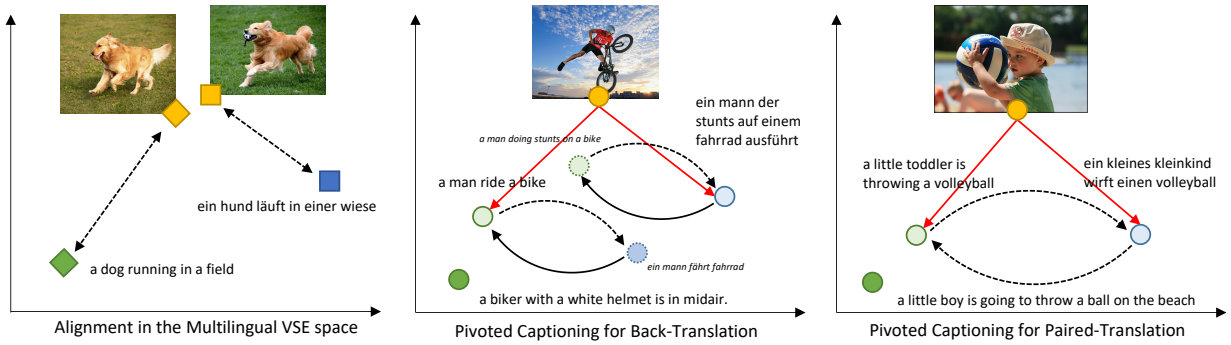


Figure 7.11: Pseudo visual pivoting: (1) multilingual VSE (src-*img*-tgt, in fact src-img$_1$, tgt-img$_2$), and (2) pivoted captioning (*src*-img-*tgt*). The *italic* items do not exist and are approximated (pseudo). (src, img, tgt) is colored in (green, yellow, blue). Solid red and black lines indicate captioning and translation without updates. Encoder-decoder are updated with dashed lines to improve the alignments in the multilingual multimodal embedding space.

multimodal MT, the multimodal back-translation objective can be extended as:

$$\begin{aligned}\mathcal{L}_{x \leftrightarrow y}^{MBT} = & \mathbb{E}_{(\mathbf{x},\mathbf{z}_x)}\Big[\text{-log } p_{yz \to x}\left(\mathbf{x}|g^*(\mathbf{x},\mathbf{z}_x),\mathbf{z}_x\right)\Big] \\ & + \mathbb{E}_{(\mathbf{y},\mathbf{z}_y)}\Big[\text{-log } p_{xz \to y}\big(\mathbf{y}|h^*(\mathbf{y},\mathbf{z}_y),\mathbf{z}_y\big)\Big]\end{aligned} \tag{7.18}$$

I simplify the notation of expectation for clarity.

7.4.4 Visual Pseudo Pivoting

Aligning the latent spaces of the source and target languages without supervision is challenging, as discussed in (Lample et al., 2018b). However, as people speak different languages biologically share similar visual systems, I envision that the shared visual space can serve as the pivot for alignment. Unlike most previous work (Chen et al., 2018b; Su et al., 2019) treating images merely as a feature, I propose two visual pivoting approaches: (1) Aligning the multilingual VSE space; (2) Image pseudo pivoting via captioning. As illustrated in Fig. 7.11, for (1), I use images as the approximate pivots connecting real non-parallel sentences. (src-*img*-tgt.) In (2), for each pivoting real image, I generate captions in both languages to construct "pseudo" source-target sentence pairs. (*src*-img-*tgt*), where the *italic* item is "pseudo". I collectively term the proposed approach *pseudo visual pivoting*.

Multilingual Visual-Semantic Embedding

I posit that for $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$, the two language spaces $\mathcal{X}, \mathcal{Y}$ could be properly associated by respectively aligning two monolingual VSE spaces $\mathcal{X} \leftrightarrow \mathcal{Z}$ and $\mathcal{Y} \leftrightarrow \mathcal{Z}$. I leverage the contrastive objective in cross-modal retrieval (Kiros et al., 2014; Huang et al., 2019c) for aligning multimodal inputs in the shared VSE space where the embeddings are close if they are semantically associated or paired.

Specifically, I generalize the fine-grained (object-level and token-level), monolingual textual-to-visual, and visual-to-textual attention (Lee et al., 2018; Huang et al., 2019d) into the multilingual setup. For fine-grained image-sentence alignment, let $s_{ij} = \cos(\mathbf{h}_i^x, \mathbf{h}_j^z)$ denotes the cosine similarity between the i-th encoded token and the j-th encoded visual object. The image-sentence similarity can be measured by averaging the cosine similarities between the visually-attend sentence embeddings and the visual embeddings of the objects. The visually-attended sentence embeddings $\mathbf{h}^{zx}$ are the weighted combination of the encoded tokens $\mathbf{h}^x$. Precisely, I compute $\mathbf{h}_j^{zx} = \sum_{i=1}^{N} \alpha_{ij} \mathbf{h}_i^x$, where $j = 1 \cdots K$ and $\alpha_{ij} = \text{softmax}_i(s_{ij})$. Let us denote by $S(\mathbf{x}, \mathbf{z}) = \frac{1}{2K} \sum_{j=1}^{K} \cos(\mathbf{h}_j^{zx}, \mathbf{h}_j^z) + \frac{1}{2N} \sum_{i=1}^{N} \cos(\mathbf{h}_i^{xz}, \mathbf{h}_i^x)$ as the image-sentence similarity, the contrastive triplet loss encouraging image-sentence alignment in the VSE space can be written as:

$$\begin{aligned} \mathcal{L}_c(\mathbf{x}, \mathbf{z}) = & \max_{\tilde{\mathbf{x}}} \big[\gamma - S(\mathbf{x}, \mathbf{z}) + S(\tilde{\mathbf{x}}, \mathbf{z})\big]_+ \\ & + \max_{\tilde{\mathbf{z}}} \big[\gamma - S(\mathbf{x}, \mathbf{z}) + S(\mathbf{x}, \tilde{\mathbf{z}})\big]_+, \end{aligned} \tag{7.19}$$

where $[.]_+$ is the hinge function, and $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{z}}$ are the non-paired (negative) instances for $\mathbf{x}$ and $\mathbf{z}$. Intuitively, when the loss decreases, the matched images and sentences will be drawn closer down to a margin γ than the hardest non-paired ones. Formally, I minimizing the following objective for cross-modal alignments in the two VSE spaces:

$$\mathcal{L}_{x,y,z}^{VSE} = \mathbb{E}_{(\mathbf{x},\mathbf{z}_x)} \Big[\mathcal{L}_c(\mathbf{x}, \mathbf{z}_x)\Big] + \mathbb{E}_{(\mathbf{y},\mathbf{z}_y)} \Big[\mathcal{L}_c(\mathbf{y}, \mathbf{z}_y)\Big] \tag{7.20}$$

Image Captioning for Pseudo Pivoting

Inspired by back-translation with monolingual corpora, I propose a novel cross-modal approach to generate automatic weakly supervised pairs to guide unsupervised MMT. Specifically, I leverage image captioning to synthesize pseudo sentence pairs (pivoted and conditioned on the image) for back-translation and paired-translation.

Image captioning models are akin to MT models besides the non-sequential visual encoder. For example, an image-to-source captioning model estimates the likelihood as $p_{z \to x}(\mathbf{x}|\mathbf{z}) =$

$\prod_{i=1}^{N} p(x_i|\mathbf{x}_{<i}, \mathbf{z})$, where $\mathbf{z}$ is the encoded image. Essentially, the captioning model learns to minimize the following loss:

$$\mathcal{L}_{z\to x}^{CAP} = \mathbb{E}_{(\mathbf{z}_x,\mathbf{x})}\left[-\log p_{z\to x}(\mathbf{x}|\mathbf{z}_x)\right] \tag{7.21}$$

As illustrated in Fig. 7.11, I incorporate two captioning models $\mathcal{Z} \to \mathcal{X}$ and $\mathcal{Z} \to \mathcal{Y}$ to generate additional "pseudo" parallel sentences pivoted on the image as additional weak supervision to better align language latent spaces in unsupervised MMT. For example, with Image $\to$ English and Image $\to$ German, the generated pseudo (English, German) pair is then pivoted on the Image. Learning captioning models is practical as it is easier to collect large-scale image-text pairs than translation pairs. I pre-train these captioning models and use them to generate sentences in two languages depicting the same image, *i.e.* $c_x^*(\mathbf{z}_x) = \text{argmax} p_{z\to x}(\mathbf{x}|\mathbf{z}_x)$ and $c_y^*(\mathbf{z}_x) = \text{argmax} p_{z\to y}(\mathbf{y}|\mathbf{z}_x)$. The pivoted captions then enable the following two objectives:

Pivoted Captioning for Back-Translation I utilize the synthetic multilingual captions (*i.e.* $c_x^*(\mathbf{z}_x)$, $c_y^*(\mathbf{z}_x)$ from the source images and $c_x^*(\mathbf{z}_y)$, $c_y^*(\mathbf{z}_y)$ from the target images) to reversely reconstruct the synthetic captions from their translations in both directions. Formally, I compute the following caption-based back-translation loss:

$$\begin{aligned}\mathcal{L}_{x\leftrightarrow y}^{CBT} = \mathbb{E}_{\mathbf{z}_x}\Big[&\text{-log } p_{yz\to x}\Big(c_x^*(\mathbf{z}_x)|g^*(c_x^*(\mathbf{z}_x),\mathbf{z}_x),\mathbf{z}_x\Big)\\ &\text{-log } p_{xz\to y}\Big(c_y^*(\mathbf{z}_x)|g^*(c_y^*(\mathbf{z}_x),\mathbf{z}_x),\mathbf{z}_x\Big)\Big]\\ +\mathbb{E}_{\mathbf{z}_y}\Big[&\text{-log } p_{yz\to x}\Big(c_x^*(\mathbf{z}_y)|h^*(c_x^*(\mathbf{z}_y),\mathbf{z}_y),\mathbf{z}_y\Big)\\ &\text{-log } p_{xz\to y}\Big(c_y^*(\mathbf{z}_y)|h^*(c_y^*(\mathbf{z}_y),\mathbf{z}_y),\mathbf{z}_y\Big)\Big]\end{aligned} \tag{7.22}$$

Pivoted Captioning for Paired-Translation With the synthetic "pseudo" paired (source, target) captions pivoted on a image (*e.g.*($c_y^*(\mathbf{z}_x)$, $c_x^*(\mathbf{z}_x)$), the caption-based paired-translation loss is defined as:

$$\begin{aligned}\mathcal{L}_{x\leftrightarrow y}^{CPT} = \mathbb{E}_{\mathbf{z}_x}&\Big[\text{-log } p_{xz\to y}(c_y^*(\mathbf{z}_x)|c_x^*(\mathbf{z}_x),\mathbf{z}_x)\Big]\\ +\mathbb{E}_{\mathbf{z}_y}&\Big[\text{-log } p_{yz\to x}(c_x^*(\mathbf{z}_y)|c_y^*(\mathbf{z}_y),\mathbf{z}_y)\Big]\end{aligned} \tag{7.23}$$

Note that similar to the text back-translation, for $\mathcal{L}_{x\leftrightarrow y}^{CPT}$ and $\mathcal{L}_{x\leftrightarrow y}^{CBT}$, I do not back-prop through the captioning step. For optimization, I sample mini-batches and minimizing the following loss:

$$\mathcal{L} = \mathcal{L}_{x\leftrightarrow y}^{MBT} + \mathcal{L}_{x,y,z}^{VSE} + \mathcal{L}_{x\leftrightarrow y}^{CBT} + \mathcal{L}_{x\leftrightarrow y}^{CPT} \tag{7.24}$$

Here I drop the weights w of each loss for clarity. In practice, all the weights are set to 1.0 except for w_{CPT} where I employ a decreasing learning scheduler specified in the next section.

7.4.5 Empirical Evaluation

Dataset and Pre-processing

I conduct experiments on the Multi30K (Elliott et al., 2016) dataset, the benchmark dataset for multimodal MT. It contains 29K training, 1K validation, and 1K testing images. Each image has three descriptions in English/German/French, which are translations of each other.

To ensure the model never learn from parallel sentences, I randomly split Multi30K training and validation sets in half for one language and use the complementary half for the other. The resulting M30k-half are two corpora with non-overlapping 14,500 training and 507 validation image-sentence pairs, respectively.

For text pre-processing, I use Moses (Koehn et al., 2007) scripts for tokenization and apply the Byte Pair Encoding (BPE) (Sennrich et al., 2016b) from XLM. To identify and extract features of visual objects in images, I use the Faster-RCNN (Ren et al., 2015) model in (Anderson et al., 2018) to detect up to 36 salient visual objects per image and extract their corresponding 2048-dim regional features.

Implementation

I use Transformer as the underlying architecture for the translation and captioning modules. Each encoder/decoder of the translator is with 6-layer stacked Transformer network, 8 heads, 1024 hidden units, and 4096 feed-forward filter size. The captioner is a 6-layer Transformer decoder with the same configuration. The visual encoder is a 1-layer MLP which maps visual feature to the shared 1,024-dim embedding space then adds the positional encoding to encode spatial locations (normalized top-left and bottom-right coordinates) of visual objects. My implementation is based on the codebase of XLM and MASS.

Experimental Details

I respectively conduct unsupervised MMT experiments on Multi30K-half for two language pairs: English-French and English-German. The followings are the pre-training, fine-tuning, and model selection protocols.

Pre-Training Pre-training is a critical step for unsupervised MT. I follow the setup in UMMT (Su et al., 2019) for a fair comparison. For each language, I create a text-only pre-training set by combining the shuffled first 10 million sentences of the WMT News Crawl datasets from 2007 to

2017 with 10 times of M30k-half, resulting in a text-only dataset with 10.145 million unparalleled sentences in English, French, German respectively.

For text pre-training, I leverage the script and the masked seq-to-seq objective proposed in MASS, which randomly masks a span in a sentence then encourages the model to decode and reconstruct the masked sequence as the monolingual language model pre-training. More details can be found in the original paper. Note that there is no fine-tuning (back-translation) on WMT for a fair comparison with other baselines.

For multimodal pre-training of the captioning modules, I use the out-of-domain MS-COCO (Lin et al., 2014) dataset. I randomly split the training set into two disjoint subsets. Each set contains 56,643 images and 283,215 sentences. I use the translate-train strategy as in XNLI (Conneau et al., 2018b). I leverage Google Translate to translate one set of English sentences into French and German. I pre-train the captioning modules with Eq. 7.21 and fix them during fine-tuning to avoid overfitting. Note that the captioning modules are trained on non-parallel sentences with disjoint image subsets, which implies no overlap between English-German or English-French sentences.

Fine-tuning on Multi30K-half I fine-tune on the training set of Multi30K-half for 18 epochs. I train the model with the Adam optimizer (Kingma and Ba, 2014) with a linear warm-up and a learning rate varying from 10^{-7} to 10^{-5}. I apply a linearly decreasing weight from 1.0 to 0.1 at 10-th epoch for w^{CPT} as I empirically observe that the generated captions are relatively too noisy to serve as good pseudo pairs in the later stage of training. The margin γ in VSE is set to 0.1. Other hyper-parameters in Transformer follow the default setting in MASS. I use 4 Titan Xp GPUs with 1,000 tokens in each mini-batch for training.

Evaluation and Model selection For evaluation, I report BLEU scores by multi-bleu.pl[4] in Moses and METEOR[5] scorea on the Multi30K testing set.

For model selection without a parallel validation corpus, I consider the unsupervised criterion proposed in (Lample et al., 2018a) based on the BLEU scores of "round-trip" translations (source $\rightarrow$ target $\rightarrow$ source and target $\rightarrow$ source $\rightarrow$ target) which have been empirically shown to correlate well with the testing metrics.

[4]https://github.com/moses-smt/mosesdecoder/blob/master-/scripts /generic/multi-bleu.perl
[5]https://github.com/cmu-mtlab/meteor

	en→fr		fr→en		en→de		de→en	
Model	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
MUSE† (Conneau et al., 2018a)	8.5	-	16.8	-	15.7	-	5.4	-
UNMT† (Lample et al., 2018a)	32.8	-	32.1	-	22.7	-	26.3	-
XLM† (Lample and Alexis, 2019)	46.3	64.3	42.0	38.1	27.4	48.7	30.7	31.0
MASS† (Song et al., 2019)	49.8	65.8	43.7	38.7	30.2	51.3	32.5	33.4
Game-MMT (Chen et al., 2018b)	-	-	-	-	16.6	-	19.6	-
UMMT-T† (Su et al., 2019)	37.2	33.7*	38.5	36.4	21.0	25.4*	25.0	28.4
UMMT-Full (Su et al., 2019)	39.8	35.5*	40.5	37.2	23.5	26.1*	26.4	29.7
Ours-Text only†	49.5	65.7	43.5	38.5	30.1	51.5	32.4	33.0
Ours-Full	**52.3**	**67.6**	**46.0**	**39.8**	**33.9**	**54.1**	**36.1**	**34.7**

Table 7.7: **Results on unsupervised MT**. Comparison with benchmarks on the Multi30K testing set. The full model is with T+V+VSE+CBT+CPT. The best score is marked bold. † means text-only. * is the METEOR score shown in the UMMT paper.

Baseline Models I compare recent unsupervised text-only and multimodal MT baselines listed in the following: (1) MUSE (Conneau et al., 2018a) is a word-to-word MT model with pre-trained Wikipedia embeddings. (2) UNMT (Lample et al., 2018a) sets the tone of using denoising autoencoder and back-translation for unsupervised MT. (3) XLM (Lample and Alexis, 2019) deploys masked language model from BERT. (4) MASS (Song et al., 2019) uses a masked seq-to-seq pre-training objective, achieves the current state-of-the-art performance in text-only unsupervised MT. (5) Game-MMT (Chen et al., 2018b) is a reinforcement learning-based unsupervised MMT. (6) UMMT (Su et al., 2019) use visual feature for denoising autoencoder and back-translation. UMMT is the current state of the art in unsupervised MMT. I either use the reported scores in the original papers or use their best scripts with their pre-trained language models publicly available for fine-tuning on Multi30K-half.

Main Results: Unsupervised MMT

Comparison with the Baseline Models Table 7.7 presents the benchmark results with other state-of-the-art unsupervised MT and MMT models on the Multi30K testing set. The first four rows show the results of the recent text-only MT models. Game-MMT and UMMT are MMT models using both image and text inputs. My full model (T+V+VSE+CBT+CPT) yields new state-of-the-art performance in BLEU and METEOR, outperforming the text-only and multimodal

baseline model by a large margin. Notably, my full model outperforms UMMT by +5.5∼12.5 BLEU scores, sets a new state of the art in unsupervised MMT.

Although pre-training plays a vital role in unsupervised MT, comparing Ours-Text only and Ours-Full, the results suggest that multimodal content can further boost the performance for unsupervised MT. Images provide +2.7∼3.7 BLEU score improvement across four tasks. Note that my model uses different monolingual pre-training corpora to MASS and XLM for the fair comparison with UMMT. With a similar pre-training objective, the text-only model is worse than MASS, while Ours-Full outperforms MASS by +2.3∼3.7 in BLEU.

Comparing the multimodal models trained with and without visual content (UMMT-T vs. UMMT-Full) and (Ours-T vs. Ours-Full), my model achieves +2.5∼3.7 improvements in BLEU while +1.4∼2.5 for UMMT. The results imply that, even with a higher text-only baseline (*e.g.*49.5 vs. 37.2 in en →fr), pseudo visual pivoting incorporates visual content more effectively.

Ablation Studies To quantify module-wise contribution in pseudo visual pivoting, I summarize the ablation studies in Table 7.8. Comparing the performance improvement from text-only to the model with regional visual features (T+V), the features of salient visual objects contribute +0.6∼0.9 BLEU score over a much higher text-only baseline compared to UMMT.

In pseudo visual pivoting, +VSE promotes the alignments in the monolingual VSE spaces and results in an additional +1.3∼2.0 gain in BLEU. This improvement validates my hypothesis that the visual space can effectively serve as the bridge connecting the source and target language latent spaces. Also, synthesizing image-pivoted pseudo caption pairs effectively provides weak supervision for aligning the cross-lingual latent space in unsupervised MMT. I observe that the pivoted captions for paired translation (CPT) is more effective than treating them as back-translation pairs (CBT). Utilizing generated image-pivoted captions is shown to be a promising approach for weakly supervised or unsupervised MMT. The full model which employs VSE, CBT, and CPT achieves +1.9∼3.1 improvements compared to the multimodal baseline (row two, visual feature only).

Generalizability How does the proposed unsupervised MMT model generalize when images are not available at the testing time? Table 7.9 shows the testing results *without* images. As can be observed, my model generalizes well. The differences are mostly less than 1.0 in BLEU. As for my model, when being tested without visual content, still outperforms other unsupervised text-only or multimodal MT models listed in Table 7.7, the minor drop in BLEU implies that the improved cross-lingual latent space alignment via pseudo visual pivoting is likely to be more

Model (Ours)	en→fr	fr→en	en→de	de→en
Text only	49.52	43.48	30.10	32.35
T+V	50.43	44.10	31.01	32.95
T+V+VSE	51.72	45.73	32.67	34.94
T+V+CPT	51.64	45.55	33.04	35.02
T+V+CBT	51.23	45.21	32.51	33.87
T+V+VSE+CBT	51.81	45.83	33.01	34.38
T+V+CPT+CBT	51.85	45.65	33.61	35.85
T+V+VSE+CPT	52.19	**46.10**	33.73	35.60
Full Model	**52.29**	45.98	**33.85**	**36.07**

Table 7.8: Ablation studies. BLEU comparison of different training objectives.

Model	en→fr	fr→en	en→de	de→en
UMMT	39.44$_{-0.35}$	40.30$_{-0.23}$	23.18$_{-0.34}$	25.47$_{-0.92}$
Ours-no VSE	51.60$_{-0.25}$	45.39$_{-0.26}$	33.25$_{-0.36}$	**35.15**$_{-0.70}$
Ours-Full	**51.64**$_{-0.65}$	**45.48**$_{-0.50}$	**33.32**$_{-0.53}$	35.04$_{-1.03}$

Table 7.9: BLEU of testing full model with text-only inputs. Subscripts are the difference to testing with T+V.

critical than using images as an input feature for decoding. Luckily, such alignment is already preserved in the training phase with the proposed approach.

An interesting question is: How much does the visual content (as a feature) contribute? As in leave-one-feature-out cross-validation, I compare the difference of performance between inferencing with and without images. The larger the difference (the subscripts in Table 7.9) implies a model better utilizes visual content. Compared with UMMT, my model has better utilization. I observe that the key to such difference is the VSE objective. The model trained without the VSE objective results in worse utilization (smaller difference at the testing time), possibly because the source text-image pairs are distant in the multilingual VSE space.

Real-pivoting & Low-resource Corpora Will my model benefit from "real" pivoting (src-img$_1$, img$_1$-tgt, overall src-img$_1$-tgt)? I train my models with overlapped images while leaving sentences in the source and target languages unparalleled (use *no* translation pairs). From the first three rows in Table 7.10, the performance is improved when training with the overlapped images and their corresponding sentences. Comparing the improvement from 0% to 100% of the

Img overlap % (# imgs/sents)	en→fr	fr→en	en→de	de→en
0% (14.5K/14.5K)	52.29	45.98	33.85	36.07
50% (22K/22K)	55.13	47.54	34.61	37.01
100% (29K/29K)	**58.34**	**50.57**	**35.45**	**38.55**
0% (T only/14.5K)	49.52	43.48	30.10	32.35
100% (T only/29K)	53.35	46.27	31.35	34.06
0% (3.0K/3.0K)	31.48	27.91	23.94	26.60
0% (T only/3.0K)	30.33	26.95	21.65	23.47

Table 7.10: Testing BLEU of the full T+V model and the text-only model trained with overlapped images or low-resource unpaired corpora.

text-only model and the full model, a larger gain is observed with the proposed pseudo visual pivoting which aligns and reduces uncertainty in the language latent spaces.

Furthermore, under the low-resource setting (3.0K non-parallel data, row six and seven), a substantial improvement over the text-only model is still observed. These results suggest that the proposed pseudo visual pivoting is likely to generalize to the semi-supervised and the low-resource setting, which I consider as the future work.

Qualitative Results

In Fig. 7.12, I provide some qualitative results on the Multi30K testing set. I observe a consistent improvement of unsupervised translation quality with the full model to the text-only one. Without parallel translation pairs as the vital supervision, the proposed pseudo visual pivoting successfully disambiguates the word semantics in the similar syntactic category and results in improved cross-lingual word alignment; for instance, "cafe" vs. "soda" machine in the third French example, and "felsigen" (rocky) vs. "verschneiten" (snowy) in the first German example.

Additional Results: Supervised MMT

Although the proposed pseudo visual pivoting targets unsupervised MMT, I are also interested in its performance under the fully supervised setup. To gain insights, I conduct supervised MMT experiments by changing the back-translation objective for unsupervised MT (Eq. 7.18) to the supervised MT objective (Eq. 3.2) with additional visual inputs. I benchmark with recent supervised MMT models, including Imagination (Elliott and Kádár, 2017), LIUM-CVC (Caglayan

T: un jeune garçon se tient sur un chariot de vêtements .
T+V: un jeune garçon s'apos accroche à un poteau de vêtements
GT: un jeune garçon s'apos accroche à un portant .
SRC: a young boy is hanging onto a clothing rack .

T: un chat assis sur le sommet d'apos un magasin de vêtements
T+V: un chat est assis sur un panneau de magasin .
GT: un chat est assis sur une enseigne de magasin .
SRC: a cat sits on top of a store sign .

T: deux garçons en train de faire une machine à café .
T+V: deux garçons devant une machine à soda .
GT: deux garçons devant une machine à soda .
SRC: two boys in front of a soda machine .

(a) English→French

T: ein mann und eine junge auf einem verschneiten strand .
T+V: ein mann und ein junge auf einem felsigen strand .
GT: ein mann und ein junge auf einem felsigen strand .
SRC: a man and a boy on a rocky beach .

T: mann springt mit einem felsbrocken im hintergrund .
T+V: mann springt vor einer felsformation im hintergrund in die luft
GT: mann springt vor einer felsformation im hintergrund .
SRC man jumping with a rock formation in background .

T: zwei männer spielen gitarre im freien .
T+V: zwei männer spielen gitarre vor einem großen publikum .
GT: zwei männer spielen gitarre vor einem großen publikum .
SRC: two men playing guitar in front of a large audience .

(b) English→German

Figure 7.12: Qualitative results of the proposed model. GT: ground truth. T+V: Full model.

	en→fr		en→de	
Model	BLEU	METEOR	BLEU	METEOR
Imagination	-	-	30.2	51.2
LIUM-CVC	52.7	69.5	30.7	52.2
VAG	53.8	70.3	31.6	52.2
Ours (T)	65.2	**79.3**	42.0	60.5
Ours (T+V)	**65.5**	79.1	**42.3**	**60.6**

Table 7.11: Supervised MMT results on Multi30K

et al., 2017), and VAG (Zhou et al., 2018c) on Multi30K.

Table 7.11 shows the testing results. My model significantly outperforms other baselines and achieves state-of-the-art performance. Comparing to the unsupervised model trained with full Multi30K (Table 7.10,100% (29K/29K)), the direct supervision from parallel translation pairs results in a +6.5∼7.1 gain in BLEU. Notably, images provide a minor improvement with full supervision from translation pairs. This result implies that, compared to serving as a complementary feature, visual information likely contributes more to improving cross-lingual alignment via pseudo visual pivoting for MMT with limited supervision.

7.5 Summary

In this chapter, I focused on fine-tuning multilingual vision-language models under limited supervision. When English-vision annotations are insufficient, I proposed to leverage automatically extracted regional semantics from un-annotated images as the additional weak supervision. My

method employs adversarial attentive alignments to alleviate the inherent heterogeneous gaps between the annotated and un-annotated portions of visual and textual data. The experimental results show that the proposed model outperforms other models by a significant margin in cross-modal retrieval tasks on the sparse Flickr30k and MS-COCO datasets. Regarding annotation efficiency, when the textual annotations are only sparsely available, my model achieves on-par performance to the recent state-of-the-art models while using up to 80% fewer annotations.

When non-English annotations are insufficient, as shown in the thesis roadmap (Fig. 1.1), it is feasible to rely on zero-shot cross-lingual transfer as discussed in section §6.3 or exploit text-only or multimodal MT models to translate-train English training data into non-English languages.

I built two multimodal MT models: supervised multimodal MT and unsupervised multimodal MT to exploit the additional visual information in vision-language tasks or datasets. For the supervised multimodal MT introduced in section §7.3, as one of the earliest works in this field, I successfully incorporated visual content for improving text-only neural machine translation. The visual content is shown to provide complementary information and the interactions of source-visual-target content in the multilingual text-image embedding space effectively improve the translation quality.

Besides the supervised multimodal MT model, I introduced my unsupervised multimodal MT empowered with a novel *pseudo visual pivoting* approach. Beyond features, the proposed method utilize visual content to improve the cross-lingual alignments in the shared latent space. Precisely, my model utilizes the visual space as the approximate pivot for promoting alignments in the shared multilingual multimodal embedding space. It synthesizes image-pivoted pseudo sentences in two languages and pairs them to translate by reconstruction without parallel corpora. The experiments on Multi30K show that the proposed model generalizes well and yields new state-of-the-art performance.

Chapter 8

Conclusion

This chapter summarizes the work presented in the thesis, highlights its major accomplishments, discusses the lessons learned, and looks forward to the future direction.

8.1 Accomplishments

8.1.1 The Journey and My Recommendation

This thesis facilitates cross-lingual generalization of vision-language models via multilingual multimodal pre-training at scale and multilingual multimodal fine-tuning under limited supervision. As depicted in Fig. 8.1, I have identified clear and practical paths towards multilingual vision-language models under different scenarios. The journey towards multilingual vision-language models is summarized as below:

Started from adversarial probing the existing image-text models, the work in Chapter §4 confirmed the advantages of leveraging attention-based models to learn object-level representations that improve text-to-image retrieval on MS-COCO and Flick30K. Extended to the multilingual scenario, the study in Chapter §5 showcased that incorporating multilingual data and promoting the diversity in multi-head attention achieved state-of-the-art multilingual text-to-image search performance on Multi30K at that time and provided multilingual entity-to-object grounding.

Scaled up to pre-train on multi-million instructional videos and their corresponding transcriptions, in Chapter §6, I alleviated the bottleneck of supporting set in contrastive learning and proposed multilingual multimodal Transformers to learn representations at scale. My work achieved state-of-the-art text-to-video retrieval on 4 datasets. Meanwhile, I constructed Multi-HowTo100M, a collection of million-scale instructional videos and their transcriptions in 9 languages. The

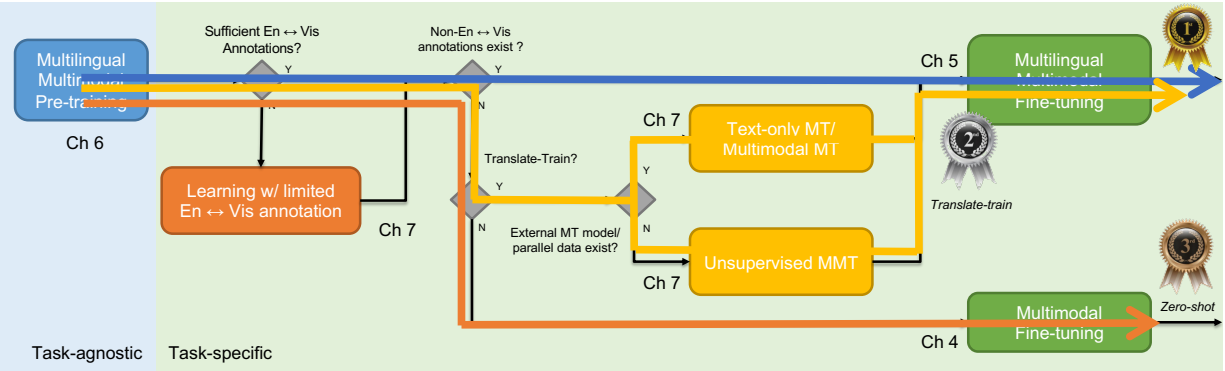


Figure 8.1: **The suggested path for cross-lingual generalization of vision-language models**. Performing multilingual multimodal pre-training is recommended. For an end task, use in-domain human-annotated non-English annotations when they are available. If they are not, use supervised/unsupervised multimodal MT to translate-train. As a case study, in Chinese-to-video search on VATEX, using in-domain human-annotated Chinese captions results in 40.5 R@1. Using supervised multimodal MT achieves 32.6 R@1. The zero-shot cross-lingual transfer (with multilingual multimodal pre-training) yields 29.7 R@1.

proposed task-agnostic multilingual multimodal pre-training effectively generalizes English-based vision-language models to handle non-English inputs without collecting additional in-domain annotations in the target language. This approach set a milestone for generalizing English-based vision-language models to multilingual vision-language models.

Finally, Chapter §7 addresses the challenges of multilingual multimodal fine-tuning on the end task where English-vision and non-English annotation are likely to be insufficient. With limited English-vision annotations, my work leverages visual semantics to synthesize additional image-semantic pairs for training where the heterogeneous domain gaps are further alleviated via adversarial learning.

To enable multilingual multimodal fine-tuning on English-only data, I introduced supervised and unsupervised multimodal MT that exploits visual information in vision-language tasks to synthesize non-English-vision data for training. For supervised multimodal MT, my study on multimodal attention identified feasible ways to incorporate complementary visual information and improved text-only MT. To further alleviate the need for parallel corpus, I introduced unsupervised multimodal MT empowered with "pseudo visual pivoting". Pivoted on the visual space, my model learns a multilingual text-image embedding space to associate text sub-spaces for aligning between source and target words. Based on the visual content, it also synthesizes additional source-image-target pairs for back-translation and paired-translation. This approach significantly improves

unsupervised multimodal MT on Multi30K.

8.1.2 Advantages of Learning Multilingual Multimodal Representations

Collectively, compared to multimodal representations, this thesis identifies several advantages of learning multilingual multimodal representations:

Robustness Learning multilingual cross-view representation results in more generalized and robust representations. It improves multiple end tasks including:

1. Multilingual text-to-video retrieval (Chapter §6)
2. Multilingual text-to-image retrieval (Chapter §5 and §6)

Grounding and Interpretability Adopting multilingualism in multimodal representations enables a new degree-of-freedom for model grounding and model interpretability, examples are:

1. Explaining multilingual image-text search: Providing explicit visual object to multilingual token alignments. (Chapter §5)
2. Visually-grounded bilingual word translation: Enabling and improving visually grounded bilingual word translation. (Chapter §5 and Chapter §7)

Complementary Multimodal Information Visual content provides complementary information that improves text-only MT in:

1. Supervised multimodal MT (Chapter §7)
2. Unsupervised multimodal MT (Chapter §7)

Transferring Linguistic Knowledge via the Visual Space Bridging via the visual space is a promising solution for pivoting that transfers linguistic knowledge more effectively. I verity this idea and demonstrate the feasibility of:

1. Cross-lingual transfer of English-based vision-language models to non-English languages (Chapter §6)
2. Unsupervised multimodal MT (Chapter §7)

8.1.3 Contributions

This thesis contributes towards promoting multilingual generalizability of vision-language models via facilitating learning multilingual multimodal representations at scale and improving the robustness when annotations are insufficient. In the following, I highlight thesis contributions in various sub-fields and tasks:

Multilingual Multimodal Representation Learning

1. Model architecture: I show that attention-based models are the desirable architectures to distill crucial information from varying length multimodal content for contrastive learning. The multilingual multimodal Transformers achieves current state-of-the-art performance in the considered end tasks.
2. Learning objective: I develop tailored contrastive objectives to effectively handle image, video, and text in different languages. Also, I design regularizers that explicitly promote diversity among multi-head attentions and alleviate heterogeneous domain gaps between real and synthetic instances.
3. Bottlenecks in contrastive learning: I analyze the ill-contrasted issue due to shared semantics in video-text representation learning. I further alleviate such side effects by promoting a self-exclusive reconstruction objective.

Multilingual Cross-Modal Retrieval

1. I develop pure transformer-based multilingual text-video encoders and remove RNN and CNN used in prior work which are harder to parallelize.
2. I introduce multilingual multimodal pre-training, which is one of the early pre-training works target on multilingual vision-language models.
3. I explore the practical scenario where annotations are insufficient and utilize visual semantics to improve the model robustness.

Multimodal Machine Translation

1. For supervised multimodal MT, my work is one of the early works that study methods to incorporate visual content. I validate the advantages of using visual content as the complementary information source for MT and outperforms text-only MT.
2. For unsupervised multimodal MT, my work on "pseudo visual pivoting" identifies a new direction of using visual content as the bridging pivot instead of the complementary information source as in supervised multimodal MT. The method also sets a new state of the art in unsupervised multimodal MT.

Cross-lingual Generalization of Vision-Language Models

1. I empirically demonstrate that vision-language models, unlike NLP models, have limited zero-shot cross-lingual transferability.
2. I construct the Multi-HowTo100M dataset, which is currently the largest multilingual text-video collection, to promote future multilingual multimodal research.
3. My model and multilingual multimodal pre-training set a milestone for generalizing English-based vision-language models to non-English languages. The suggested roadmap is shown in Fig. 8.1.

8.2 Lessons Learned

Based on the experience gained along my research journey in large-scale multilingual multimodal representation learning, in this section, I summarize the lessons learned and provide a guideline for future endeavors in this direction.

1. The trade-off in computation: There is a trade-off between computation and performance. Training a model end-to-end from raw image or video as inputs may achieve better performance. Also, incorporating longer transcriptions, or longer video clips may also achieve better performance. However, the computation and memory consumption may greatly slow down the training, especially for the video part. Understand such trade-offs and make reasonable decisions (*e.g.*, freeze the video encoder) could be crucial for related multimodal research.
2. Noise in data: There is inevitable noise in the multimodal datasets even for manually annotated MS-COCO, VATEX, Multi30K. The situation is even worse for the transcriptions in video-text datasets such as HowTo100M. It could be tempting to filter out or correct noisy instances by performing pre-processing such as lemmatization for text or stabilization/color correction for video/image to clean up the data. However, heavy pre-processing steps may only provide a marginal gain and in many cases, they are even harmful.
3. Keep complexity in mind: Although it's tempting to develop large and complicated models that may achieve better performance. However, the complexity may prevent them from being useful and deployed. For example, inter-modal attention as analyzed in section §3.2.2 usually performs better than intra-modal attention. However, its complexity is $O(MN)$ compared to $O(M + N)$ for models with intra-modal attention. In practice, that difference

is 3 hours vs 5 seconds for searching 1000 images with 1000 text queries.

4. Difference between modalities: Data in different modalities share very different characteristics and require different treatment. Visual embeddings are typically more diverse textual embeddings (in the visualization, compared to the textual embeddings, the visual embeddings usually span over the entire space). Additionally, for the contrastive objectives, triplet loss works better for image-text while NCE works better for video-text in my work. Although many alternative contrastive objectives have been proposed in recent years, they highly depend on the setup and dataset and no one prevails universally.
5. Curse of Scaling: The scale of data ambiguities the progress made in the small subset. Also, a fancy method that works well on clean small data may not apply to large-scale noisy data. In many cases, simple methods overfit less and work better on large-scale noisy data.
6. Comparing apples to apples: With the rapid development in the CV and NLP communities, some of the reported results fall short to be the apple-to-apple comparison. For example, in the field of text-to-video retrieval, different feature sets, training dataset size, and testing split make it harder to compare fairly. It's advised to be careful when reproducing the results and experiment with one's own model under the same protocol.

In sum, given that the sheer volume of multimodal data, it is advised to start from understanding and analyzing the data to be modeled. The spirit is to preserve and utilize the unique characteristics of data in different modalities and focus on innovating towards versatile models that embed sufficient capacity and feasible complexity. Although handling large-scale noisy data is challenging, one should not be discouraged. For task-agnostic multilingual multimodal pre-training, a small improvement could be translated into improvements in various end tasks.

8.3 To Jump Start

In this section, I provide my suggestion for future researchers who are interested in this important and challenging research direction to quickly jump start. My recommendations for the dataset, end task, and the corresponding codebase for pre-training and fine-tuning are as follows.

Regarding pre-training, task-agnostic multilingual multimodal pre-training at scale is shown to be an important step. One may start with the multi-HowTo100M dataset and the repository at https://github.com/berniebear/multi-ht100m. I provide the script for extracting video features, the pre-training script on Multi-HowTo100M, and the fine-tuning script on various end tasks. One may either use the pre-trained checkpoints and the model provided in the repository

or re-train on Multi-HowTo100M from scratch with one's own model. Considering the scale of the data, it will be more efficient to start with the pre-trained model provided in the repository. Note that it takes around 8TB to store the instructional videos in Multi-HowTo100M or HowTo100M. Pre-training on 9 languages takes around one day over an eight-V100 GPU instance for 16 epochs.

Besides video-text data, a possible enhancement is to consider pre-training with additional image-text data. For large-scale image-English data such as conceptual caption (Sharma et al., 2018), one may consider using multimodal MT models proposed in this thesis or off-the-shelf text-only MT models such as https://marian-nmt.github.io/ to translate English captions into non-English languages for image-text pre-training.

For fine-tuning, the codebase released focuses on multilingual text-video retrieval in VTT[1] and VATEX, as well as multilingual text-image retrieval in Multi30K. As all the model architectures are based on Transformer, other vision-language tasks such as VQA and vision-language navigation can also be covered by applying tailored classification heads for the end task. An example can be found in the codebase provided by (Xu et al., 2021). In all cases, it is desirable to collect a sufficient amount of in-domain English-vision data for training.

Collecting multilingual training data is a critical step towards multilingual vision-language models. In general, in-domain human annotations are the most reliable. Alternatively, one may leverage multimodal or text-only MT models to translate the training data into the target language(s) and train the model. Supervised multimodal or text-only MT models are more robust in comparison to the unsupervised ones when a large-scale parallel corpus is available. For high resource languages where monolingual corpora are available, exploiting unsupervised multimodal MT will outperform the zero-shot cross-lingual transfer. It's important to note that training with multiple language-vision data is typically better than training with monolingual data.

8.4 Looking Forward

For future work, I discussed two directions that have great potential yet are currently under-explored.

[1]Note that the evaluation protocols (such as the training and testing splits used in VTT) vary significantly from one paper to another.

8.4.1 Multilingual Multimodal Self-Supervised Learning

There are 3 modalities covered in this thesis include image, video, and text in multiple languages. For each modality, the instances share notable variances. The videos scene and length of MSR-VTT differ significantly from Multi-HowTo100M. Also, the text types are very different in MS-COCO (*i.e.*, captions) and instructional videos in HowTo100M (*i.e.*, transcriptions of the demonstrators' speech). There are many possible modalities to be included for cross-view self-supervised learning, such as audio and depth. Looking forward, unifying and covering more modalities and richer instances types are considered as reasonable next steps.

In the short term, developing approaches that unify information from multiple modalities to perform multi-dataset training is a reasonable next step. One may additionally introduce audio, or combine image-text and video-text training by sharing the visual encoder where an additional temporal pooling layer could be applied to videos. To handle different types of text annotation, similar to the generative augmentation discussed in Chapter §7, generators or captioning modules could be pre-trained on each dataset and be applied for synthesizing additional textual data (could be in different languages) on the visual data of other datasets.

Learning and modeling multimodal human knowledge would be the long-term goal. Current progress in this thesis is mainly built on large-scale data and self-supervised methods such as contrastive learning to associate object-level or entity-level knowledge elements in different modalities. However, human intelligence goes beyond such simple alignments within or across modalities. For example, relationships between textual and visual objects and hierarchies of knowledge elements go beyond the scope of this thesis. To this end, investigating capacious models and effective algorithms that unify various learning paradigms (supervised, self-supervised, and semi-supervised) and task objectives (discriminative, generative task, and so on) would be the immediate future research direction.

8.4.2 Towards Multilingual Multimodal Vision-Language Models

The work presented in this thesis has identified a practical path towards generalized multilingual multimodal vision-language models. The direct next step could be resolving the performance gap between zero-shot and training with in-domain non-English annotations, to achieve a better cross-lingual transfer performance on par with the transfer performance in NLP. Beyond improving system performance, there are remaining two challenges to be resolved: (1) language coverage and (2) task coverage.

For language coverage, the work in task-agnostic multilingual multimodal pre-training (Chap-

ter §6) covers 9 languages, which is only a small portion of the existing 6,500 $\sim$ 7000 languages. Also, the user-generated and machine-translated captions are very limited in size. For example, the MT models utilizing Wikipedia[2] corpus, only cover 109 languages. For some low-resource languages, there are no reliable MT models nor user-annotated captions. Additionally, the underlying "visual pivoting" assumption presented in this thesis may not generalize well to more abstract concept. People from different countries may have very different social norms that uniquely shape the way they speak and describe. Also, there are many culture-dependent concepts where "visual pivoting" may fail. For example, there is no "carnival" in the Chinese culture. To this end, developing systematic approaches to increase the language coverage in multilingual multimodal vision-language models, especially in low-resource domains, is an open question.

Instead of pivoting on shared videos, one feasible approach could be exploiting data from multiple multimodal datasets as discussed before. For example, utilizing both multilingual image-text (*e.g.*, Multi30K and YJ-captions-26k (Miyazaki and Shimizu, 2016)) and multilingual video-text datasets (*e.g.*, Multi-HowTo100M and VATEX) to integrate the existing annotation efforts. The other approach is to jointly model and train translation models and multimodal contrastive learning under a multi-task setup, which may achieve better performance when training them separately. Also, as in mBERT, a more sophisticated language type sampling may improve the multilingual multimodal pre-training in Chapter §6.

For vision-language task coverage, this thesis mainly covers the multilingual cross-modal retrieval tasks and translation tasks. Developing effective architectures and techniques to transfer varieties of tasks including (1) discriminative vision-language models (*e.g.*, VQA (Goyal et al., 2017), TVQA (Lei et al., 2020), V-L Navigation), and (2) generative vision-language models (*e.g.*, multilingual captioning, multilingual text-to-image synthesis) and (3) visually-enhanced NLP models (*e.g.*, unsupervised multimodal MT (Huang et al., 2020b)) are the promising future directions to be explored.

[2]https://en.wikipedia.org

Appendix A

Appendix

A.1 Additional details for Multilingual Multimodal Pre-training

In this section I provide additional details for multilingual multimodal pre-training and the Multi-HowTo100M dataset described in §6.3. This section is organized as follows: First I provide details about the Multilingual HowTo100M (Multi-HowTo100M) dataset for multilingual multimodal pre-training (MMP) in §A.1.1. Then I provide additional implementation details and experiment setup in §A.1.2. Additional ablation studies regarding choices of Transformer architecture are discussed in §A.1.3. Then I present additional cross-dataset transfer experiments in §A.1.4.

A.1.1 The Multilingual HowTo100M Dataset

In this section I provide the detailed statistics of the Multilingual HowTo100M (Multi-HowTo100M) dataset. I also provide a comparison to Sigurdsson et al. (2020) that also uses HowTo100M for unsupervised word translation.

The Multi-HowTo100M dataset is built upon the original English HowTo100M dataset Miech et al. (2019) that contains 1.2 million instructional videos (138 million clips) on YouTube. I reuse the *raw* English subtitles in HowTo100M, where the subtitles in HowTo100M are either automatic speech recognition (ASR) transcriptions or user generated subtitles.

For Multi-HowTo100M, I use the same video collection as English HowTo100M. At the time of data collection (May 2020), there were 1.09 million videos accessible. I collect the subtitles provided by YouTube, which either consist of user-generated subtitles or those generated by Google ASR and Translate in the absence of user-generated ones. Essentially, I collect video subtitles in 9 languages: English (*en*), German (*de*), French (*fr*), Russian (*ru*), Spanish (*es*), Czech

Language	videos	#subtitle	#tokens
English	1238911	138429877	1.18B
German	1092947	69317890	1.26B
French	1093070	69399097	1.33B
Czech	1092717	68911940	1.22B
Russian	1092802	69117193	1.25B
Chinese	1092915	68939488	0.94B
Swahili	1092302	68898800	1.22B
Vietnamese	1092603	68887868	1.13B
Spanish	1092649	70143503	1.16B

Table A.1: Multi-HowTo100M statistics

(*cz*), Swahili (*sw*), Chinese (*zh*), Vietnamese (*vi*). Table A.1 summarizes the dataset statistics for each language. In most cases there are more than 1 billion tokens a language.

Fig. A.1 further shows the number of tokens per video. There are typically lengthy narrations that contains several hundreds of tokens available in each instructional video. Fig. A.2 shows the distribution of number of tokens in a subtitle. For each subtitle segment, which ranges from 0$\sim$20 seconds, there are typically 15$\sim$25 words. The most of the cases, subtitles are well aligned in time for non-English languages. Fig. 6.5 visualizes a few examples in Multi-HowTo100M.

A similar HowTo100M variant has been recently reported in MUVE (Sigurdsson et al., 2020) that is created for unsupervised word translation. Our Multi-HowTo100M differs from MUVE in the following perspectives: First, I collects 9 language for *all* videos in HowTo100M while MUVE only has 4 languages available (English, French, Japanese, and Korean) on HowTo100M. Also, MUVE divided HowTo100M into 4 non-overlapped sections for each language, there are no parallel pairs for each subtitle. While in Multi-HowTo100M, there are 7-9 languages for each subtitle. Essentially, There are more than 1 billion tokens in most languages in Multi-HowTo100M. To our best knowledge, our Multi-HowTo100M dataset is currently the largest multilingual text-video collection.

Beyond scale, instructional videos in Multi-HowTo100M are feasible pre-training resources for many downstream vision-language models. Demonstrators in instructional videos typically perform intentionally and explain the visual object or action explicitly. According to the inspection by (Miech et al., 2019), for around 51% of clips, at least one object or action mention in the caption can be visually seen. Prior work has shown that instructional videos are useful for

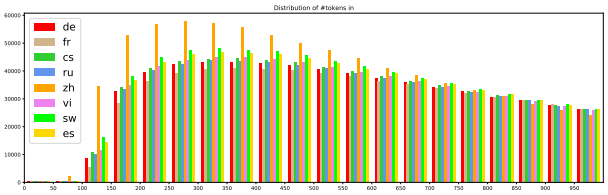


Figure A.1: Distribution of #tokens/video in Multi-HowTo100M

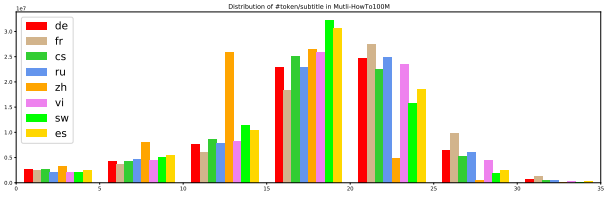


Figure A.2: Distribution of #tokens/subtitle in Multi-HowTo100M

event recognition Yu et al. (2014), action localization model Alayrac et al. (2016), cross-modal alignments Malmaud et al. (2015). I expect the previous success in the intersection of natural language processing (NLP) and computer vision (CV) could be further translated into more languages to have a broaden impact.

The are great potentials of using our Multi-HowTo100M dataset in related research field such as multilingual multimodal representation learning Huang et al. (2019b); Kim et al. (2020); Burns et al. (2020), multilingual multimodal translation Huang et al. (2020b); Surís et al. (2020), multilingual image/video captioning Miyazaki and Shimizu (2016) ... etc. I expect the release of Multi-HowTo100M will be a first step towards spurring more research in these directions.

A.1.2 Implementation and Experiment Details

Pre-Processing. For pre-possessing, I truncate the maximum length N of text to 192 for pre-training on Multi-HowTo100M. The maximum length is set to 96 for fine-tuning VTT Xu et al. (2016), VATEX Wang et al. (2019) and Multi30K Elliott et al. (2016). The maximum video length M is set to 128 for pre-training on Multi-HowTo100M and 36 for all fine-tuning tasks.

Model Architecture. For the multilingual Transformers, either multilingual BERT Devlin et al. (2019b) or XLM-R-large Artetxe et al. (2020), I use the pre-trained version provided by HuggingFace. [1] and use their corresponding tokenizers for tokenization. Detailed design choices regarding output layer and frozen layer is discussed in §A.1.3.

[1] https://github.com/huggingface/transformers

For the video backbone, I use a 34-layer, R(2+1)-D (Tran et al., 2018) network pre-trained on IG65M Ghadiyaram et al. (2019) and a S3D Miech et al. (2020) network pre-trained on HowTo100M Miech et al. (2019). I apply a spatial-temporal average pooling over the last convolutional layer, resulting in a 512-dimensional vector for each 3D CNN network. I extract visual features at a rate of 1 feature per second. Since the 3D CNNs employs different size of input windows (*e.g.*, 8 frames for R(2+1)D and 16 for S3D), I re-sample videos to 30 fps and employs a window of size 8 or 30 that takes consecutive frames starting from the beginning of every second for encoding. I simply concatenate the two 3D-CNN outputs and use the 1024-dimension vector as the visual input stream to our model. Notably, instead of using 9 different types of visual features as in CE (Liu et al., 2019), I use only the above 2 features and achieve superior performance.

For the Transformer pooling head (TP) modules, I use a 2-layer Transformer with 4-head attention for each TP. The embedding dimension D is set to 1024. I do not use the positional embeddings in both text and video TP as I do not find them beneficial in our experiments. The softmax temperature in all NCE contrastive objectives is set to 0.1 as used in SimCLR Chen et al. (2020b).

Note that unlike ViLBERT Lu et al. (2019) or OAN Huang et al. (2019d), our models does not employ cross-modality attention and keep the multi-head self-attention within the same modality. The main reason is to reduce the inference time complexity. For cross-modality attention, the complexity is $O(TV)$ to encode T text queries for V videos in a dataset before retrieval (since video and query representations depend on each other). It is clearly not scalable when the dataset contains millions of videos. To this end, our model keep self-attention within the same modality which results in a $O(T+V)$ complexity compared $O(TV)$ in prior work with cross-modality attention. In our preliminary experiments, I also incorporate cross-modality attention and achieved 0.3$\sim$1.8 R@1 improvement. Considering the trade-off between performance and scalability, I choose the latter.

Training and Inference Details and Profiling. For the softmax temperature in NCE, I set to 0.1 as used in SimCLR Chen et al. (2020b). I use the Adam (Kingma and Ba, 2014) optimizer with a initial learning rate $2 \cdot 10^{-4}$ and clip gradients greater than 0.2 during the training phase. Dropout rate is 0.3. Since the video length and token length is longer in the pre-training phase, I use a 64 batch size for pre-training. For fine-tuning, I use a batch size of 128.

Pre-training on the 1.2 million HowTo100M videos takes around 10 GPU hours (NVIDA V100) for 16 epochs. I speed up the pre-training process by distributing the workload over 8 GPUs on a single node of our server. I use 1 GPU for the fine-tuning or training from scratch experiments.

For the MSR-VTT split, it takes 12 GPU hours to train our model on 180K video-text pairs for 20 epochs. For VATEX, it takes 32 GPU hours to train on 260K video-text pairs for 30 epochs. For inference, the encoding speed is around 250-300 videos/sec and 200-250 text queries/sec. The overall text→video search speed on 1,000 video-text pairs (1,000 text queries over 1,000 videos) is around 6 seconds including video/text encoding and ranking their similarity scores.

Experiment Details. Our experiment consider three types of pre-training: (1) Multilingual multimodal pre-training (MMP), (2) Multimodal pre-training (MP), and (3) no pre-training (from scratch). For (1) and (2), I pre-train 16 epochs and use the model weight at 16-th epoch for fine-tuning experiments.

For multimodal pre-training, I pre-train on the original English HowTo100M dataset. I iterate over all videos in HowTo100M. For each video, I randomly sample the start and end time to construct a video clip. For each clip, I locate the nearest consecutive ASR transcriptions in time and use it as to construct the (video, text) pair for training.

For multilingual multimodal pre-training (MMP), I use Multi-HowTo100M for pre-training. For each video, I follow the same strategy as MP. For a clip, I sample one language type each time from 9 languages and use the consecutive ASR transcriptions that are closest in time to compose (video, text) pairs for training.

After pre-training, I fine-tune our model on VTT and VATEX to evaluate on text→video search tasks. In the zero-shot cross-lingual transfer experiments, I use only English-video data. I then directly test the model with non-English queries to report the zero-shot performance. When annotations in additional languages are available (by humans in VATEX and Multi30K; by MT models (*i.e*. *translate-train*) in VTT), I train our model with all available multilingual annotations (*i.e*. fully supervised) to compare fairly with other baselines in multilingual text→video search.

Since pre-trained model has a faster convergence rate, I fine-tune for 10 epochs and use the model with best validation performance (summation of R@1, R@5, R@10) for testing. For models without pre-training (*i.e*., from-scratch), I train for 20 epochs under the same training protocol.

A.1.3 Additional Ablation Studies

As has been investigated in XTREME Hu et al. (2020), choosing different output layers will affect the zero-shot transferability of multilingual Transformers in various NLP tasks. For text→video search tasks, I conduct a series of experiments to identify the desirable choices of hyper-parameters

Output layer	Freeze lower	*en*	*de*
3	0	20.9	3.2
6	0	20.5	3.1
9	0	21.0	4.8
12	0	21.0	13.3
15	0	20.5	12.3
18	0	20.8	12.6
12	6	21.0	15.5
12	9	**21.0**	**16.3**
12	12	18.9	14.1

Table A.2: Text→video R@1 of XLM-R output layers and layers to freeze on VTT

Output layer	Freeze lower	*en*	*de*
3	0	19.2	2.5
6	0	19.5	2.0
9	0	19.3	5.8
12	0	19.6	8.8
12	6	19.3	10.5
12	9	**19.9**	**11.1**
12	12	18.9	9.8

Table A.3: Text→video R@1 of mBERT output layers and layers to freeze on VTT

in the proposed multilingual multimodal Transformer that lead to best performance in English-to-video and (zero-shot) non-English-to-video search performance. Beyond our ablation studies in Sec. 5, in this part I highlight our trials in the choice of the output layer and the layers to be frozen in our multilingual Transformer backbone (*i.e.*, mBERT and XLM-R). There are 24 layers in XLM-R (large) and 12 layers in mBERT. I perform grid-search on VTT to identify the best choice of these two hyper-parameters.

Choice of Output Layers Table A.2 and Table A.3 compare different choices of output layer and layers to freeze in multilingual Transformers. Our results suggest that the best output layer for mBERT and XLM-R is the 12-th layer. Surprisingly, while output layer does not affect English→video search significantly, it greatly affects the zero-shot cross-lingual transfer performance of video-text models. For both XLM-R and mBERT, the performance degrade

text→video	English	Non-English
In-domain	✓	✓
Out-of-domain	✓	

Table A.4: Coverage of our experiments

significantly if fine-tuning all layers.

Choice of Layers to Freeze Similar to output layers, the choice of frozen layers greatly affects cross-lingual transferability. For both mBERT and XLM-R, it is desirable to freeze part of the lower layers and make the top-3 layers trainable for video-text models. I observe that when freezing all layers (*i.e.*, using the pre-extracted contextual multilingual embeddings) does not lead to satisfactory results. For mBERT, R@1 drops from 19.9 to 18.9 in English→video search and 11.1 to 9.8 in German→video search. For XLM-R, R@1 drops from 21.0 to 18.9 in English→video search and 16.3 to 14.1 in German→video search. These results imply that text-only contextual multilingual embeddings along are likely to be infeasible to be applied to vision-language tasks without proper fine-tuning.

An important observation is that the best English→video search performance corresponds to the best German→video performance. This trend implies that for model selection, the configuration for the best English→video model usually translates to the best configuration for (zero-shot) cross-lingual model. This shared trend justifies the English→video ablation studies in the original paper. Note that I utilize the best English→video for all (zero-shot) cross-lingual experiment in our experiment section.

For multilingual text→video search, the best configuration I found in our experiments is to output the 12-th layer and freeze the layers below 9 for both mBERT and XLM-R.

A.1.4 Additional Experimental Results

The coverage of our text→video search experiments is summarized in Table A.4. Our experiments cover the following scenarios:

In-domain, English: Table 5 (VTT) and Table 6 (VATEX) in the original paper.

In-domain, non-English: Table 4 (VTT, 9 languages) and Table 6 (VATEX, Chinese).

Out-of-domain, English: Additional (zero-shot) generalization results across datasets are in §A.1.5.

Out-of-domain, non-English: I consider this as our future work.

Model	R@1	R@5	R@10
VSE (Kiros et al., 2014)	10.1	29.4	41.5
VSE++ (Faghri et al., 2018)	14.4	35.7	46.9
Dual (He et al., 2016a)	13.7	36.1	48.2
HGR (Chen et al., 2020a)	16.4	38.3	49.8
Ours-Full	**24.0**	**50.5**	**62.1**

Table A.5: Zero-shot generalization on YouTube2Text with VTT-finetuned model.

A.1.5 Generalizability across English-Video Datasets

In this section. I provide additional experiment results regarding zero-shot generalization of the VTT-finetuned model on out-of-domain dataset. Specifically, I test on YouTube2Text Guadarrama et al. (2013). The aim of this experiment is to test the cross-dataset generalizabilty of our model without using domain-specific training data.

Table A.5 shows the comparison of English→video search results on the YouTube2Text testing set. Models in this table are only fine-tuned on VTT and use *no* YouTube2Text training data. As can be observed, our model with MMP generalizes well on YouTube2Text, outperforming HGR Chen et al. (2020a) by 7.6 and DualEncoder He et al. (2016a) by 10.3 in R@1.

A.2 Additional details for Bottlenecks in Multimodal Representation Learning

This appendix provides additional details for analyzing bottlenecks in multimodal representation learning §6.2. First, I provide more details about the model. Then I introduce the datasets and the experimental setup. Finally, I provide additional qualitative and quantitative experimental results for text-video retrieval and captioning.

A.2.1 Model Details

Implementation details and hyper parameters. For our text encoder, I use the T5-base model pre-trained on the "Colossal Clean Crawled Corpus" (C4) (Raffel et al., 2019). I use its corresponding text tokenizer and encode a sentence into a sequence of 1024 dimensional vectors.

For our visual encoder, our model utilizes only the motion and the appearance features. For the motion feature, I use a 34-layer, R(2+1)-D (Tran et al., 2018) model pre-trained on

IG65M (Ghadiyaram et al., 2019) and apply a spatial-temporal average pooling over the last convolutonal layer, resulting in a 512-dimensional vector. For the appearance feature, I use the 2048-dimension flattened pool-5 layer of the standard ResNet152 (He et al., 2016b) pre-trained on Imagenet (Deng et al., 2009). I extract features at a rate of 1 feature per second and simply concatenate the two features, resulting in a 2560-dimension visual input stream. Noteworthily, instead of using 9 and 7 different types of visual features as in CE (Liu et al., 2019) and MMT (Gabeur et al., 2020), I use only the above 2 features and achieve on par or superior performance. Also, with early fusion, our model does not suffer from additional computation required for the extended sequence length in MMT. For the text decoder, I use the T5-base model decoder, also pre-trained on C4.

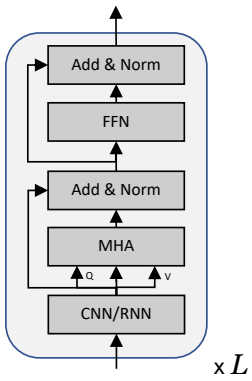


Figure A.3: Transformer pooling head.

As illustrated in Fig. A.3, our transformer pooling head is composed of a pre-encoder, a multi-head self-attention (MHA), and a feed-forward layer (FFN). For pre-encoders, I use a one-layer MLP with a d-dimensional output for mapping video features into the common embedding space. I use 1024-dimension bi-directional GRU as the text pre-encoder. For the 1D-CNN prior, I use kernels with size $[2, 3, 4, 6]$ as the visual and text pre-encoders. I set the embedding dimension to 1024 and use 4 attention heads in the transformer pooling layers. The hidden dimension of FFN is 2048.

Training and Inference time. Pre-training on 1.2 million HowTo100M videos takes around 160 GPU hours (NVIDIA V100) for 20 epochs. I speed up the pre-training process by distributing the workload over 8 GPUs. I use 1 GPU for the fine-tuning or training from scratch experiments. For the MSR-VTT 1k-A split, it takes 12 GPU hours to train our full model on 180K video-text pairs for 20 epochs. For Vatex, it takes 32 GPU hours to train on 260K video-text pairs for 30 epochs. For ActivityNet, it takes 2.5 GPU hours to train on 10K video-text paris for 28 epochs.

For inference, the encoding speed is around 250-300 video/sec and 200-250 text query/sec. The overall text-to-video search speed on 5,000 video-text pairs (5,000 text queries over 5,000 videos) is 30-34 seconds including encoding. The speed of text-to-video retrieval is similar to video-to-text retrieval.

A.2.2 Experiment Details

The margin α of the max-margin loss is 0.2, and the temperature T is set to 0.1 as used in SimCLR Chen et al. (2020b). I use the Adam (Kingma and Ba, 2014) optimizer with a initial learning rate $5 \cdot 10^{-5}$ and clip gradients greater than 0.2 during the training phase. Dropout rate is 0.3 for all datasets besides ActivityNet (0.0).

As the average video/text lengths and videos available are quite different across datasets, I adjust our training scheme accordingly. When training on MSR-VTT, ActivtyNet and Vatex, batch-size is set to 64. For MSR-VTT training, I sample and truncate videos to 32 seconds, text to 100 tokens and train for 20 epochs. For Vatex, videos are at most 64 seconds and I train for 30 epochs. For ActivtityNet training, videos are at most 512 seconds and 256 tokens for the text part. I train for 28 epochs on ActivityNet. For fine-tuning HowTo100M pre-trained model, I reduce training epochs into quarters.

A.2.3 Video Captioning Experiments

To measure captioning/text generation performance, I report BLEU4 (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2014), Rogue-L (Lin, 2004) and CIDEr (Vedantam et al., 2015) metrics. I report results on the MSR-VTT, VATEX and ActivityNet datasets.

Table A.6: **Captioning performance on the MSR-VTT dataset**

	Captioning			
	BLUE4	METEOR	Rogue-L	CIDEr
VidTranslate (Korbar et al., 2020)	41.7	28.5	–	–
POS+VCT (Hou et al., 2019)	42.3	29.7	62.8	49.1
ORG (Zhang et al., 2020b)	43.6	28.8	62.1	50.9
Ours, MSR-VTT only	39.7	28.3	60.5	46.5
Ours, HT100M + MSR-VTT	38.9	28.2	59.8	48.6

Table A.7: **Captioning performance on the VATEX dataset**

	Captioning			
	Blue@4	METEOR	Rogue-L	CIDEr
Shared Enc-Dec (Wang et al., 2019)	28.4	21.7	47.0	45.1
ORG (Zhang et al., 2020b)	32.1	22.2	48.9	49.7
Ours, VATEX only	**32.8**	**24.4**	**49.1**	**51.2**
Ours, HT100M + Vatex	32.5	24.1	48.9	50.5

Table A.8: **Captioning performance on the ActivtyNet dataset**

	Captioning			
	Blue@4	METEOR	Rogue-L	CIDEr
DENSE (Krishna et al., 2017a)	1.6	8.9	–	–
DVC-D-A (Li et al., 2018)	1.7	9.3	–	–
Bi-LSTM+TempoAttn (Zhou et al., 2018b)	2.1	10.0	–	–
Masked Transformer (Zhou et al., 2018b)	2.8	11.1	–	–
Ours, ActivityNet only	1.5	6.9	17.8	3.2
Ours, HT100M + ActivityNet	1.4	6.9	17.5	3.1

A.2.4 Zero-Shot Retrieval Experiments

I also evaluate our model in the zero-shot setting on MSR-VTT, Vatex, ActivityNet and MSVD, after pre-training on HT100M. While I are able to get reasonable results on MSR-VTT and MSVD, our results are not great on Vatex and Activity-Net due to significant domain gap.

A.2.5 Action Recognition Experiments

Lastly, I evaluate our model on the video action recognition task on HMDB-51 (Kuehne et al., 2011) and UCF-101 (Soomro et al., 2012). For this, I use the R(2+1)D-34 (pretrained on IG65M) model as well as a ResNet-152 model (pretrained on Imagenet), as in our method. I extract a feature per second per video by concatenating the features from each model (2560-D), and obtain an average representation per video using either average pooling (2560-D) or our proposed transformer pooling head (1024-D) pre-trained on HT100M using cross-captioning objective. I then train a linear classifier for 1500 epochs for HMDB-51 (500 for UCF-101) on these features using Adam (Kingma and Ba, 2014) optimizer with learning rate of $1e^{-4}$ and weight decay $1e^{-4}$

Table A.9: **Zero-shot Retrieval performance on VATEX, MSR-VTT, MSVD and ActivityNet.**

	Text →Video				Video →Text			
	$R@1\uparrow$	$R@5\uparrow$	$R@10\uparrow$	$MdR\downarrow$	$R@1\uparrow$	$R@5\uparrow$	$R@10\uparrow$	$MdR\downarrow$
Zero-Shot								
ActivityNet	0.06	0.2	0.5	1907.0	0.0	0.2	0.3	2238.0
VATEX	0.07	0.4	0.7	682.0	0.07	0.4	0.9	697
MSVD	8.9	26.0	37.9	18.0	21.4	46.2	57.7	6.0
MSR-VTT	8.7	23.0	31.1	31.0	12.7	27.5	36.2	24.0

with early stopping. I also drop the learning rate by 10 at epochs 200, 400 for UCF-101 and 1000, 1200 for HMDB-51. In Table A.10, I show the results of training only a linear-layer on features extracted from our fixed backbone with or without a learned transformer-pooling head. I find that our transformer temporal pooling head provides significant benefits over the baseline of simply average pooling the features, demonstrating the effectiveness of building contextualized representations using our proposed transformer. In particular, I see improvements of over 7% on HMDB-51 and 34% on UCF-101 by replacing average pooling with our transformer pooling head to aggregate features. I observe that naive average pooling performs significantly worse than our transformer pooling under evaluation protocol. This is likely because 1) the average pooling collapses temporal information, making the linear layer based classification difficult 2) compared to the transformer pooling, it does not benefit from large-scale pretraining on a wide variety of action videos of HT100M. I further compare very favorably to the current state-of-the-art approaches. In particular, I outperform all other approaches, both supervised and self-supervised, except the recently introduced Omni (Duan et al., 2020) which was finetuned on both UCF-101 and HMDB-51, while I only trained a linear classifier on extracted features. However, it should be noted that it is very difficult to fairly compare all these different approaches because they may use different modalities (images, RGB video, optical flow, audio, ASR outputs), pretraining datasets (Kinetics-400, HT100M, IG65M, Imagenet), architectures (S3D, I3D, R(2+1)D, R3D), pre-training (supervised, self-supervised) and downstream training (frozen, finetuned) strategies.

A.2.6 Statistical significance

In Table A.11, I show the results of finetuning our pretrained model for 3 times on the VATEX dataset. I find that the variance is quite low and our model consistently beats the state of the art.

A.2.7 Additional Qualitative Results

I provide addition qualitative text-to-video retrieval results on MSR-VTT, VATEX, ActivityNet in Fig. A.4. Given a text query, in most cases, our model successfully retrieves the correct videos marked in green.

Table A.10: **Action recognition.** Results of training only a linear-layer, on features extracted from our fixed backbone with or without a learned transformer-pooling head. We compare to the state-of-art supervised and self-supervised pretrainig methods on the HMDB-51 and UCF-101 action recognition task, for different downstream training protocols ("FT?" stands for finetuned). We report average Top-1 accuracy across all 3 folds. Dataset abbreviations: **A**udio**S**et, **HMDB51**, **H**owTo100**M**, **I**nsta**g**ram**65M**, **IM**agenet-1000, **K**inetics**400**, **O**mni**S**ource Images + Videos, **S**ports**1M**, **UCF101**, **YouTube8M**. Other abbreviations: **V**ideo modality, **F**low modality, **I**mage modality, **A**udio modality, **T**ransformer pooling, **Av**era**g**e pooling

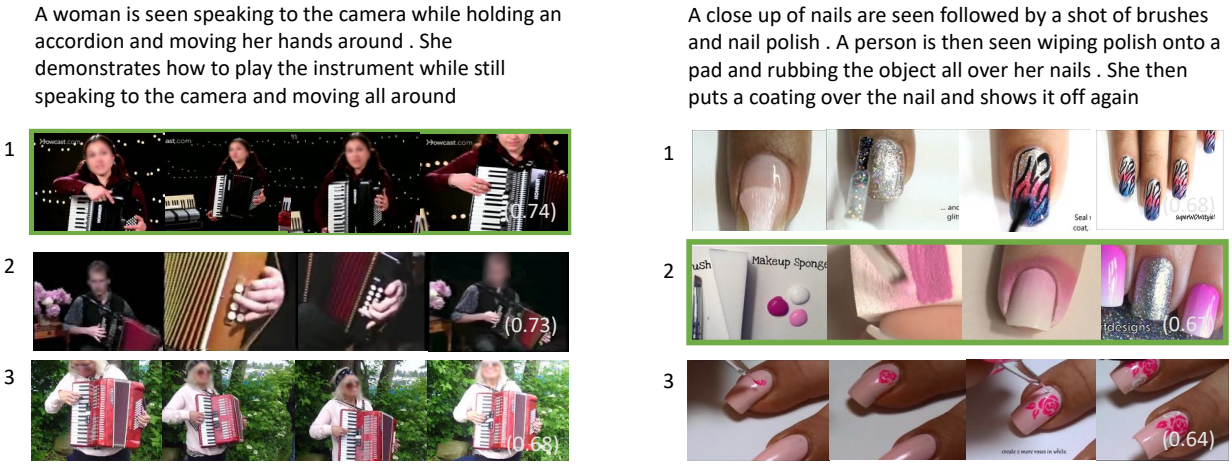
Method	Mod	Dataset	Model	FT?	H51	U101
Self-Supervised Pre-training						
MIL-NCE (Miech et al., 2020)	V,T	HM	S3D-G	✗	53.1	82.7
MIL-NCE (Miech et al., 2020)	V,T	HM	S3D-G	✓	61.0	91.3
MMV (Alayrac et al., 2020)	V,T,A	HM+AS	TSM-50x2	✗	67.1	91.8
ELo (Piergiovanni et al., 2020)	V,F,A	YT8M	R(2+1)D-50x3	✓	67.4	93.8
XDC (Alwassel et al., 2020)	V,A	IG65M	R(2+1)D-18	✓	68.9	95.5
Supervised Pre-training						
P3D (Qiu et al., 2017)	V,I	S1M+IM	P3D	✓	–	88.6
TSN (Wang et al., 2018a)	V,I	IM	TSN	✓	69.4	94.2
I3D (Carreira and Zisserman, 2017)	V,I	K400+IM	I3D	✓	74.8	95.6
R(2+1)D (Tran et al., 2018)	V	K400	R(2+1)D-34	✓	74.5	96.8
S3D-G (Xie et al., 2018)	V,I	K400+IM	S3D-G	✓	75.9	96.8
I3D (Carreira and Zisserman, 2017)	V,I	K400+IM	I3D	✓	77.1	96.7
R(2+1)D (Tran et al., 2018)	V	K400	R(2+1)D-34	✓	76.4	95.5
R(2+1)D (Tran et al., 2018)	V,F	K400	R(2+1)D-34x2	✓	78.7	97.3
Omni (Duan et al., 2020)	V,I	K400+OS	Slow-8x8-R101	✓	79.0	97.3
I3D (Carreira and Zisserman, 2017)	V,F,I	K400+IM	I3Dx2	✓	80.7	98.0
Omni (Duan et al., 2020)	V,F,I	K400+OS	Slow-8x8-R101x2	✓	**83.8**	**98.6**
Ours (Avg-pooling)	V,I	IG65M+IM	R(2+1)D-34+R152	✗	73.7	64.3
Ours (T-pooling)	V,I	HM+IG65M+IM	R(2+1)D-34+R152	✗	81.3	98.0

Table A.11: **Retrieval performance on the VATEX dataset**

	Text →Video				**Video →Text**			
	R@1↑	R@5↑	R@10↑	MdR↓	R@1↑	R@5↑	R@10↑	MdR↓
Random Baseline	0.2	0.7	1.05	2000.5	0.02	0.1	1.02	2100.5
VSE (Kiros et al., 2014)	28.0	64.3	76.9	3.0	–	–	–	–
VSE++ (Faghri et al., 2018)	33.7	70.1	81.0	2.0	–	–	–	–
Dual (He et al., 2016a)	31.1	67.4	78.9	3.0	–	–	–	–
HGR (Chen et al., 2020a)	35.1	73.5	83.5	2.0	–	–	–	–
Ours	**44.9**$_{\pm 0.2}$	**82.1**$_{\pm 0.2}$	**89.7**$_{\pm 0.2}$	**1.0**	**58.4**$_{\pm 0.1}$	**84.4**$_{\pm 0.2}$	**91.0**$_{\pm 0.3}$	**1.0**

(a) MSR-VTT

(b) VATEX



(c) ActivityNet

Figure A.4: Examples of top-3 Text→Video retrieval results and similarities on the MSR-VTT, VATEX, and ActivityNet testing set. Only one correct video (colored in green) for each text query on the top.

Bibliography

Waleed Abdulla. Mask r-cnn for object detection and instance segmentation on keras and tensorflow. https://github.com/matterport/Mask_RCNN, 2017. 84

Jean-Baptiste Alayrac, Piotr Bojanowski, Nishant Agrawal, Ivan Laptev, Josef Sivic, and Simon Lacoste-Julien. Unsupervised learning from narrated instruction videos. In *Computer Vision and Pattern Recognition (CVPR)*, 2016. 133

Jean-Baptiste Alayrac, A. Recasens, Rosália G. Schneider, R. Arandjelović, Jason Ramapuram, J. Fauw, Lucas Smaira, S. Dieleman, and Andrew Zisserman. Self-supervised multimodal versatile networks. *ArXiv*, abs/2006.16228, 2020. 55, 144

Humam Alwassel, Bruno Korbar, Dhruv Mahajan, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. In *NeurIPS*, 2020. 144

Elad Amrani, Rami Ben-Ari, Daniel Rotman, and Alex Bronstein. Noise estimation using density estimation for self-supervised multimodal learning. *arXiv preprint arXiv:2003.03186*, 2020. 63, 77

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018. 1, 14, 16, 27, 29, 33, 46, 84, 91, 112

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015. 1

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. Unsupervised neural machine translation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018. URL https://openreview.net/forum?id=Sy2ogebAW. 24, 105

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.421. URL https://www.aclweb.org/anthology/2020.acl-main.421. 66, 67, 72, 133

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1409.0473. 9, 23, 30, 56, 99

Jeremy Barnes, Lilja Øvrelid, and Erik Velldal. Sentiment analysis is not solved! assessing and probing sentiment classification. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 12–23, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4802. URL https://www.aclweb.org/anthology/W19-4802. 67

Andrea Burns, Donghyun Kim, Derry Wijaya, Kate Saenko, and Bryan A. Plummer. Learning to scale multilingual representations for vision-language tasks. In *The European Conference on Computer Vision (ECCV)*, 2020. 69, 79, 133

Ozan Caglayan, Walid Aransa, Adrien Bardet, Mercedes Garcia-Martinez, Fethi Bougares, Loïc Barrault, Marc Masana, Luis Herranz, and Joost Van de Weijer. Lium-cvc submissions for wmt17 multimodal translation task. In *SECOND CONFERENCE ON MACHINE TRANSLATION*, volume 2, pages 432–439, 2017. 117

Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. Probing the need for visual context in multimodal machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4159–4170, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1422. URL https://www.aclweb.org/anthology/N19-1422. 106

Iacer Calixto and Qun Liu. Incorporating global visual features into attention-based neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 992–1003, 2017. 106

Iacer Calixto, Qun Liu, and Nick Campbell. Multilingual multi-modal embeddings for natural language processing. *arXiv preprint arXiv:1702.01101*, 2017. 4, 41

M. Caron, I. Misra, J. Mairal, Priya Goyal, P. Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020. 55

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 15, 144

David Chen and William Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 190–200, 2011. 60

Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. Fine-grained video-text retrieval with hierarchical graph reasoning. In *CVPR*, 2020a. 15, 60, 62, 64, 72, 78, 138, 145

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020b. 54, 69, 134, 140

Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger. Adversarial deep averaging networks for cross-lingual sentiment classification. *Transactions of the Association for Computational Linguistics*, 6:557–570, 2018a. 66

Yun Chen, Yang Liu, Yong Cheng, and Victor O.K. Li. A teacher-student framework for zero-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1925–1935, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1176. URL https://www.aclweb.org/anthology/P17-1176. 105, 106

Yun Chen, Yang Liu, and Victor OK Li. Zero-resource neural machine translation with multi-agent communication game. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018b. 105, 107, 108, 109, 114

Yong Cheng, Qian Yang, Yang Liu, Maosong Sun, and Wei Xu. Joint training for pivot-based neural machine translation. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 3974–3980, 2017. doi: 10.24963/ijcai.2017/555. URL https://doi.org/10.24963/ijcai.2017/555. 105, 106

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014. 15

Volkan Cirik, Louis-Philippe Morency, and Taylor Berg-Kirkpatrick. Visual referring expression recognition: What do systems actually learn? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 781–787. Association for Computational Linguistics, 2018. URL http://aclweb.org/anthology/N18-2123. 28

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In *International Conference on Learning Representations (ICLR)*, 2018a. 105, 114

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. XNLI: evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2475–2485, 2018b. URL https://www.aclweb.org/anthology/D18-1269/. 76, 113

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://www.aclweb.org/anthology/2020.acl-main.747. 4

Ryan Cotterell and Georg Heigold. Cross-lingual character-level neural morphological tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 748–759, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1078. URL https://www.aclweb.org/anthology/D17-1078. 66

Bo Dai and Dahua Lin. Contrastive learning for image captioning. In *Advances in Neural Information Processing Systems*, pages 898–907, 2017. 1, 27

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 2, 139

Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380, 2014. 25, 140

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep

bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019a. URL https://www.aclweb.org/anthology/N19-1423/. 2, 3, 24, 46, 66, 68, 72

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019b. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423. 133

Ali Diba, Mohsen Fayyaz, Vivek Sharma, Amir Hossein Karami, Mohammad Mahdi Arzani, Rahman Yousefzadeh, and Luc Van Gool. Temporal 3d convnets: New architecture and transfer learning for video classification. *arXiv preprint arXiv:1711.08200*, 2017. 15

Haodong Duan, Yue Zhao, Yuanjun Xiong, Wentao Liu, and Dahua Lin. Omni-sourced webly-supervised learning for video recognition. In *ECCV*, 2020. 142, 144

N D'souza Rhett, Huang Po-Yao, and Yeh Fang-Cheng. Structural analysis and optimization of convolutional neural networks with a small sample size. *Scientific Reports (Nature Publisher Group)*, 10(1), 2020. 82

Aviv Eisenschtat and Lior Wolf. Linking image and text with 2-way nets. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 1855–1865. IEEE, 2017. 14, 37, 38

Desmond Elliott and Ákos Kádár. Imagination improves multimodal translation. In Greg Kondrak and Taro Watanabe, editors, *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers*, pages 130–141. Asian Federation of Natural Language Processing, 2017. URL https://www.aclweb.org/anthology/I17-1014/. 98, 117

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74. Association for Computational Linguistics, 2016. doi: 10.18653/v1/W16-3210. URL http://www.aclweb.org/anthology/W16-3210. 2, 4, 9, 18, 20, 21, 72, 98, 103, 112, 133

Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018. URL https://github.com/fartashf/vsepp. xvii, 1, 14, 17, 27, 28, 29, 31, 32, 37, 38, 44, 47, 57, 64, 78, 79, 83, 87, 92, 93, 95, 138, 145

Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. Bradford Books, 1998. 91

Orhan Firat, Baskaran Sankaran, Yaser Al-onaizan, Fatos T. Yarman Vural, and Kyunghyun Cho. Zero-resource translation with multi-lingual neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 268–277, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1026. URL https://www.aclweb.org/anthology/D16-1026. 105

Andrea Frome, Gregory S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 2121–2129, 2013. 15

Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *ECCV*, 2020. 55, 58, 63, 64, 139

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, pages 1180–1189, 2015. 88, 89, 91

Spandana Gella, Rico Sennrich, Frank Keller, and Mirella Lapata. Image pivoting for learning multilingual multimodal representations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2839–2845. Association for Computational Linguistics, 2017. doi: 10.18653/v1/D17-1303. URL http://aclweb.org/anthology/D17-1303. 4, 41, 42, 43, 46, 47, 68, 69, 79

Deepti Ghadiyaram, Du Tran, and Dhruv Mahajan. Large-scale weakly-supervised pre-training for video action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12046–12055, 2019. 15, 55, 69, 72, 134, 139

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *ICLR*, 2018. 54

Yunchao Gong, Liwei Wang, Micah Hodosh, Julia Hockenmaier, and Svetlana Lazebnik. Improving image-sentence embeddings using large weakly annotated photo collections. In *European conference on computer vision*, pages 529–545. Springer, 2014. 85

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 27, 80, 129

Jiuxiang Gu, Jianfei Cai, Shafiq R. Joty, Li Niu, and Gang Wang. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 37, 38

Sergio Guadarrama, Niveda Krishnamoorthy, Girish Malkarnenkar, Subhashini Venugopalan, Raymond J. Mooney, Trevor Darrell, and Kate Saenko. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*, pages 2712–2719. IEEE Computer Society, 2013. doi: 10.1109/ICCV.2013.337. URL https://doi.org/10.1109/ICCV.2013.337. 20, 138

Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 297–304, 2010. 54, 69

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. Dual learning for machine translation. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 820–828, 2016a. URL http://papers.nips.cc/paper/6469-dual-learning-for-machine-translation. 15, 59, 64, 78, 138, 145

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016b. 14, 29, 33, 39, 44, 46, 58, 108, 139

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with temporal language. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2018. 55

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997. 15

Jingyi Hou, Xinxiao Wu, Wentian Zhao, Jiebo Luo, and Yunde Jia. Joint syntax representation learning and visual cue translation for video captioning. In *ICCV*, 2019. 140

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual

generalization. *CoRR*, abs/2003.11080, 2020. 53, 66, 67, 72, 76, 135

Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. *arXiv preprint arXiv:1909.00964*, 2019a. 66

Po-Yao Huang, Frederick Liu, Sz-Rung Shiang, Jean Oh, and Chris Dyer. Attention-based multimodal neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, volume 2, pages 639–645, 2016. 8, 9, 98, 99, 106

Po-Yao Huang, Junwei Liang, Jean-Baptiste Lamare, and Alexander G. Hauptmann. Multimodal filtering of social media for temporal monitoring and event analysis. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, ICMR '18, page 450–457, New York, NY, USA, 2018a. Association for Computing Machinery. ISBN 9781450350464. 27, 28

Po-Yao Huang, Junwei Liang, Vaibhav Vaibhav, Xiaojun Chang, and Alexander Hauptmann. Informedia@ trecvid 2018: Ad-hoc video search with discrete and continuous representations. In *TRECVID Proceedings*, volume 70, 2018b. 16, 27

Po-Yao Huang, Xiaojun Chang, and Alexander Hauptmann. Multi-head attention with diversity for learning grounded multilingual multimodal representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1461–1467, Hong Kong, China, November 2019b. Association for Computational Linguistics. doi: 10.18653/v1/D19-1154. URL https://www.aclweb.org/anthology/D19-1154. 5, 7, 8, 9, 59, 68, 69, 79, 133

Po-Yao Huang, Guoliang Kang, Wenhe Liu, Xiaojun Chang, and Alexander G. Hauptmann. Annotation efficient cross-modal retrieval with adversarial attentive alignment. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, page 1758–1767, New York, NY, USA, 2019c. Association for Computing Machinery. ISBN 9781450368896. doi: 10.1145/3343031.3350894. URL https://doi.org/10.1145/3343031.3350894. 2, 8, 14, 15, 69, 110

Po-Yao Huang, Vaibhav, Xiaojun Chang, and Alexander G. Hauptmann. Improving what cross-modal retrieval models learn through object-oriented inter- and intra-modal attention networks. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, ICMR '19, pages 244–252, New York, NY, USA, 2019d. ACM. ISBN 978-1-4503-6765-3. doi: 10.1145/3323873.3325043. URL http://doi.acm.org/10.1145/3323873.3325043. 1,

5, 7, 14, 27, 28, 82, 110, 134

Po-Yao Huang, Xiaojun Chang, Alexander G. Hauptmann, and Eduard Hovy. Forward and backward multimodal nmt for improved monolingual and multilingual cross-modal retrieval. In *Proceedings of the 2020 on International Conference on Multimedia Retrieval*, ICMR '20, pages 244–252, New York, NY, USA, 2020a. ACM. ISBN 978-1-4503-7087-5/20/06. doi: 10.1145/3372278.3390674. URL http://doi.acm.org/10.1145/3372278.3390674. 79

Po-Yao Huang, Junjie Hu, Xiaojun Chang, and Alexander Hauptmann. Unsupervised multimodal neural machine translation with pseudo visual pivoting. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8226–8237, Online, July 2020b. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.731. URL https://www.aclweb.org/anthology/2020.acl-main.731. 8, 9, 80, 129, 133

Po-Yao Huang, Mandela Patrick, Junjie Hu, Graham Neubig, Florian Metze, and Alexander Hauptmann. Multilingual multimodal pre-training for zero-shot cross-lingual transfer of vision-language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2443–2459, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.195. URL https://www.aclweb.org/anthology/2021.naacl-main.195. 7, 8, 9

Poyao Huang, Ye Yuan, Zhen-Zhong Lan, Lu Jiang, and Alexander G. Hauptmann. Video representation learning and latent concept mining for large-scale multi-label video classification. *CoRR*, abs/1707.01408, 2017a. URL http://arxiv.org/abs/1707.01408. 1, 27

Yan Huang, Wei Wang, and Liang Wang. Instance-aware image and sentence matching with selective multimodal lstm. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7254–7262. IEEE, 2017b. 14, 16, 37, 38

Yan Huang, Qi Wu, and Liang Wang. Learning semantic concepts and order for image and sentence matching. *arXiv preprint arXiv:1712.02036*, 2017c. 14, 28, 37, 38

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 59

Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1): 221–231, 2012. 15

Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, 2017. 28

Huajie Jiang, Ruiping Wang, Shiguang Shan, and Xilin Chen. Learning class prototypes via structure alignment for zero-shot recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 118–134, 2018. 85

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351, 2017. doi: 10.1162/tacl_a_00065. URL https://www.aclweb.org/anthology/Q17-1024. 105

Ákos Kádár, Desmond Elliott, Marc-Alexandre Côté, Grzegorz Chrupała, and Afra Alishahi. Lessons learned in multilingual grounded language learning. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 402–412, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/K18-1039. URL https://www.aclweb.org/anthology/K18-1039. 4, 41, 42

Lukasz Kaiser and Samy Bengio. Can active memory replace attention? In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 3774–3782, 2016. 23

Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. In *NeurIPS*, 2020. 55

Nal Kalchbrenner and Phil Blunsom. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1700–1709, 2013. 104

Nal Kalchbrenner, Lasse Espeholt, Karen Simonyan, Aaron van den Oord, Alex Graves, and Koray Kavukcuoglu. Neural machine translation in linear time. *arXiv preprint arXiv:1610.10099*, 2016. 23

Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015. 1, 14, 16, 19, 20, 27, 37, 38, 39, 82

Andrej Karpathy, Armand Joulin, and Li F Fei-Fei. Deep fragment embeddings for bidirectional

image sentence mapping. In *Advances in neural information processing systems*, pages 1889–1897, 2014a. 16

Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014b. 91

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning, 2020. 54

Donghyun Kim, Kuniaki Saito, Kate Saenko, Stan Sclaroff, and Bryan A. Plummer. MULE: Multimodal Universal Language Embedding. In *AAAI Conference on Artificial Intelligence*, 2020. 68, 69, 79, 133

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 46, 58, 73, 113, 134, 140, 141

Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *NIPS Workshop*, 2014. 1, 16, 27, 47, 64, 78, 83, 110, 138, 145

Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. Associating neural word embeddings with deep image representations using fisher vectors. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4437–4446, 2015. 28

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. Inducing crosslingual distributed representations of words. In *Proceedings of COLING 2012*, pages 1459–1474, 2012. 66

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics, 2007. 112

Bruno Korbar, F. Petroni, Rohit Girdhar, and L. Torresani. Video understanding as machine translation. *ArXiv*, abs/2006.07203, 2020. 55, 62, 63, 77, 140

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *CVPR*, 2017a. 21, 55, 60, 141

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie

Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017b. 1, 14, 20, 46, 79, 91

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 14, 39

H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In *ICCV*, 2011. 141

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv preprint arXiv:1811.00982*, 2018. 91

Guillaume Lample and Conneau Alexis. Cross-lingual language model pretraining. *CoRR*, abs/1901.07291, 2019. URL http://arxiv.org/abs/1901.07291. 24, 106, 114

Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 7059–7069. Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/8928-cross-lingual-language-model-pretraining.pdf. 66, 70, 76

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. In *Proceedings of NAACL-HLT*, pages 260–270, 2016. 30

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations (ICLR)*, 2018a. 24, 105, 106, 113, 114

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049, Brussels, Belgium, October 2018b. Association for Computational Linguistics. doi: 10.18653/v1/D18-1549. URL https://www.aclweb.org/anthology/D18-1549. 24, 105, 106, 109

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 59

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. *arXiv preprint arXiv:1803.08024*, 2018. xvii, 1, 14, 16, 22, 27, 28, 30, 31, 37, 38, 39, 47, 83, 92, 93, 94, 95, 110

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. Tvqa: Localized, compositional video question answering. In *EMNLP*, pages 1369–1379, 2018. 55

Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. TVQA+: Spatio-temporal grounding for video question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8211–8225, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.730. URL https://www.aclweb.org/anthology/2020.acl-main.730. 80, 129

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.653. URL https://www.aclweb.org/anthology/2020.acl-main.653. 67

Juncheng B Li, Kaixin Ma, Shuhui Qu, Po-Yao Huang, and Florian Metze. Audio-visual event recognition through the lens of adversary. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 616–620, 2021. doi: 10.1109/ICASSP39728.2021.9415065. 27

Junnan Li, Pan Zhou, Caiming Xiong, Richard Socher, and Steven C. H. Hoi. Prototypical contrastive learning of unsupervised representations, 2020. 55

Yehao Li, Ting Yao, Yingwei Pan, Hongyang Chao, and Tao Mei. Jointly localizing and describing events for dense video captioning. In *CVPR*, 2018. 141

Yuncheng Li, Yale Song, Liangliang Cao, Joel R. Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. TGIF: A new dataset and benchmark on animated GIF description. *CoRR*, abs/1604.02748, 2016. URL http://arxiv.org/abs/1604.02748. 20

Junwei Liang, Desai Fan, Han Lu, Po-Yao Huang, Jia Chen, Lu Jiang, and A. Hauptmann. An event reconstruction tool for conflict monitoring using social media. In *AAAI*, 2017. 27

Jindrich Libovický and Jindrich Helcl. Attention strategies for multi-source sequence-to-sequence learning. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*, pages 196–202. Association for Computational Linguistics,

2017. doi: 10.18653/v1/P17-2031. URL https://doi.org/10.18653/v1/P17-2031. 108

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 140

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1, 2, 20, 46, 82, 90, 113

Zhouhan Lin, Minwei Feng, Cícero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017. URL https://openreview.net/forum?id=BJC_jUqxe. 24

Wenhe Liu, Guoliang Kang, Po-Yao Huang, Xiaojun Chang, Yijun Qian, Junwei Liang, Liangke Gui, Jing Wen, and Peng Chen. Argus: Efficient activity detection system for extended video analysis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, March 2020. 27

Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. In *BMVC*, 2019. 21, 55, 60, 63, 64, 69, 72, 77, 78, 134, 139

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 13–23, 2019. URL http://papers.nips.cc/paper/8297-vilbert-pretraining-task-agnostic-visiolinguistic-representations-for-visi
2, 22, 67, 134

Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation, 2020. 55, 63

Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal, September

2015a. Association for Computational Linguistics. doi: 10.18653/v1/D15-1166. URL https://www.aclweb.org/anthology/D15-1166. 99

Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, 2015b. 103

Lin Ma, Zhengdong Lu, Lifeng Shang, and Hang Li. Multimodal convolutional neural networks for matching image and sentence. In *Proceedings of the IEEE international conference on computer vision*, pages 2623–2631, 2015. 32

Jonathan Malmaud, Jonathan Huang, Vivek Rathod, Nicholas Johnston, Andrew Rabinovich, and Kevin Murphy. What's cookin'? interpreting cooking videos using text, speech and vision. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 143–152, Denver, Colorado, May 2015. Association for Computational Linguistics. doi: 10.3115/v1/N15-1015. URL https://www.aclweb.org/anthology/N15-1015. 133

Pascal Mettes, Dennis C. Koelma, and Cees G.M. Snoek. The imagenet shuffle: Reorganized pre-training for video event detection. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, ICMR '16, pages 175–182, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4359-6. doi: 10.1145/2911996.2912036. URL http://doi.acm.org/10.1145/2911996.2912036. 91

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2630–2640, 2019. 5, 21, 53, 55, 59, 63, 67, 71, 72, 77, 78, 131, 132, 134

Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-End Learning of Visual Representations from Uncurated Instructional Videos. In *CVPR*, 2020. 2, 5, 55, 68, 69, 72, 134, 144

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013. 15

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018. 46

Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *CVPR*, 2020. 54

Niluthpol C Mithun, Juncheng Li, Florian Metze, and Amit K Roy-Chowdhury. Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In *ICMR*. ACM, 2018a. 64

Niluthpol Chowdhury Mithun, Rameswar Panda, Evangelos E Papalexakis, and Amit K Roy-Chowdhury. Webly supervised joint embedding for cross-modal image-text retrieval. In *2018 ACM Multimedia Conference on Multimedia Conference*, pages 1856–1864. ACM, 2018b. 83

Takashi Miyazaki and Nobuyuki Shimizu. Cross-lingual image caption generation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1780–1790. Association for Computational Linguistics, 2016. doi: 10.18653/v1/P16-1168. URL http://aclweb.org/anthology/P16-1168. 129, 133

Hideki Nakayama and Noriki Nishida. Zero-resource machine translation by multimodal encoder–decoder network with multimedia pivot. *Machine Translation*, 31(1-2):49–64, 2017. 107, 108

Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual attention networks for multimodal reasoning and matching. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 2156–2164. IEEE, 2017. 1, 16, 23, 28, 33, 34, 36, 37, 47, 83, 92, 93, 95

Graham Neubig and Junjie Hu. Rapid adaptation of neural machine translation to new languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 875–880, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1103. URL https://www.aclweb.org/anthology/D18-1103. 66

Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Hierarchical multimodal lstm for dense visual-semantic embedding. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 1899–1907. IEEE, 2017. 32

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002. 25, 140

Mandela Patrick, Yuki Markus Asano, Bernie Huang, Ishan Misra, Florian Metze, João F. Henriques, and Andrea Vedaldi. Space-time crop & attend: Improving cross-modal video repre-

sentation learning. *CoRR*, abs/2103.10211, 2021a. URL https://arxiv.org/abs/2103.10211. 67

Mandela Patrick, Po-Yao Huang, Yuki M. Asano, Florian Metze, Alexander Hauptmann, João Henriques, and Andrea Vedaldi. Support-set bottlenecks for video-text representation learning. In *International Conference on Learning Representations (ICLR)*, 2021b. 7

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. URL http://www.aclweb.org/anthology/D14-1162. 15, 91

Florent Perronnin and Christopher Dance. Fisher kernels on visual vocabularies for image categorization. In *2007 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2007. 15

Eduard Hovy PI, Taylor Berg-Kirkpatrick, Jaime Carbonell, Hans Chalupsky, Anatole Gershman, Alex Hauptmann, Florian Metze, Teruko Mitamura, Aditi Chaudhary, Xianyang Chen, et al. Opera: Operations-oriented probabilistic extraction, reasoning, and analysis. In *TAC*, 2018. 1, 27

AJ Piergiovanni, Anelia Angelova, and Michael S. Ryoo. Evolving losses for unsupervised video representation learning. In *CVPR*, 2020. 144

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2641–2649. IEEE Computer Society, 2015. doi: 10.1109/ICCV.2015.303. URL https://doi.org/10.1109/ICCV.2015.303. 19

Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *proceedings of the IEEE International Conference on Computer Vision*, pages 5533–5541, 2017. 144

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019. 55, 58, 138

Afshin Rahimi, Yuan Li, and Trevor Cohn. Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1015. URL https://www.aclweb.org/anthology/P19-1015. 67

Janarthanan Rajendran, Mitesh M Khapra, Sarath Chandar, and Balaraman Ravindran. Bridge correlational neural networks for multilingual multimodal representation learning. In *Proceedings of NAACL-HLT*, pages 171–181, 2016. 4, 41, 42

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 14, 31, 33, 44, 46, 79, 84, 86, 98, 102, 103, 108, 112

Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1179–1195, 2017. doi: 10.1109/CVPR.2017.131. URL https://doi.org/10.1109/CVPR.2017.131. 1, 27

Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Chris Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. Movie description. *International Journal of Computer Vision*, 2017. 55

Andrew Rouditchenko, Angie Boggust, David Harwath, Dhiraj Joshi, Samuel Thomas, Kartik Audhkhasi, Rogerio Feris, Brian Kingsbury, Michael Picheny, Antonio Torralba, et al. Avlnet: Learning audio-visual language representations from instructional videos. *arXiv preprint arXiv:2006.09199*, 2020. 55, 63

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1599–1613, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1162. URL https://www.aclweb.org/anthology/N19-1162. 67

Holger Schwenk and Xian Li. A corpus for multilingual document classification in eight languages. In Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, May 2018. European Language Resources Association (ELRA). ISBN 979-10-95546-00-9. 67

Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation

models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 86–96, 2016a. 24, 105

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1715–1725, 2016b. 112

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*, 2018. 127

Gunnar A. Sigurdsson, Jean-Baptiste Alayrac, Aida Nematzadeh, Lucas Smaira, Mateusz Malinowski, João Carreira, Phil Blunsom, and Andrew Zisserman. Visual grounding in video for unsupervised word translation. In *CVPR*, 2020. URL https://arxiv.org/abs/2003.05078. 9, 66, 131, 132

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 14, 39, 100, 103

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mass: Masked sequence to sequence pre-training for language generation. In *International Conference on Machine Learning*, pages 5926–5936, 2019. 24, 114

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human action classes from videos in the wild. In *CRCV-TR-12-01*, 2012. 141

Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation, WMT 2016, colocated with ACL 2016, August 11-12, Berlin, Germany*, pages 543–553, 2016. URL https://www.aclweb.org/anthology/W16-2346/. 99, 106, 107

Yuanhang Su, Kai Fan, Nguyen Bach, C-C Jay Kuo, and Fei Huang. Unsupervised multi-modal neural machine translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10482–10491, 2019. 105, 107, 108, 109, 112, 114

Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. 2017. 55

Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. Learning video representations

using contrastive bidirectional transformer, 2019a. 55

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *ICCV*, 2019b. 55

Dídac Surís, Dave Epstein, and Carl Vondrick. Globetrotter: Unsupervised multilingual translation from visual alignment. *arXiv preprint arXiv:2012.04631*, 2020. 133

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112, 2014a. 23, 104

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *NeurIps*, pages 3104–3112, 2014b. 98

Oscar Täckström, Ryan McDonald, and Jakob Uszkoreit. Cross-lingual word clusters for direct transfer of linguistic structure. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 477–487, Montréal, Canada, June 2012. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N12-1052. 66

Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 55

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 15

Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 15, 58, 68, 72, 134, 138, 144

Yao-Hung Hubert Tsai, Liang-Kang Huang, and Ruslan Salakhutdinov. Learning robust visual-semantic embeddings. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3591–3600. IEEE, 2017. 85

Aäron van den Oord, Y. Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748, 2018. 62, 69

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017. 16, 23, 24, 43, 57, 59, 68

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 140

Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. *arXiv preprint arXiv:1511.06361*, 2015. 1, 14, 38, 47, 79, 83

Subhashini Venugopalan, Marcus Rohrbach, Jeff Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence – video to text. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015. 21, 55, 60

Kaiye Wang, Qiyue Yin, Wei Wang, Shu Wu, and Liang Wang. A comprehensive survey on cross-modal retrieval. *arXiv preprint arXiv:1607.06215*, 2016a. 1

Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. *IEEE transactions on pattern analysis and machine intelligence*, 41(11):2740–2755, 2018a. 144

Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, pages 5005–5013. IEEE, 2016b. 16, 28, 37

Liwei Wang, Yin Li, Jing Huang, and Svetlana Lazebnik. Learning two-branch neural networks for image-text matching tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018b. 28, 37

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 1, 2, 4, 20, 22, 60, 72, 133, 141

Michael Wray, Diane Larlus, Gabriela Csurka, and Dima Damen. Fine-grained action retrieval through multiple parts-of-speech embeddings. In *ICCV*, 2019. 63, 77

Shijie Wu and Mark Dredze. Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. *arXiv preprint arXiv:1904.09077*, 2019. 66

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang

Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144, 2016. URL http://arxiv.org/abs/1609.08144. 104

Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018. 54

Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 305–321, 2018. 144

Hu Xu, Gargi Ghosh, Po-Yao Huang, Prahal Arora, Masoumeh Aminzadeh, Christoph Feichtenhofer, Florian Metze, and Luke Zettlemoyer. Vlm: Task-agnostic video-language model pre-training for video understanding, 2021. 127

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A large video description dataset for bridging video and language. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 5288–5296. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.571. URL https://doi.org/10.1109/CVPR.2016.571. 1, 20, 21, 55, 59, 72, 133

Ran Xu, Caiming Xiong, Wei Chen, and Jason J. Corso. Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, page 2346–2352. AAAI Press, 2015. ISBN 0262511290. 21, 60

I. Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. Billion-scale semi-supervised learning for image classification, 2019. 55, 69

Fei Yan and Krystian Mikolajczyk. Deep correlation for matching images and text. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3441–3450, 2015. 14, 37

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 1, 19, 20, 21,

29, 72, 82, 90

Shoou-I Yu, Lu Jiang, and Alexander Hauptmann. Instructional videos for unsupervised harvesting and learning of action examples. In *Proceedings of the 22nd ACM International Conference on Multimedia*, MM '14, page 825–828, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450330633. doi: 10.1145/2647868.2654997. URL https://doi.org/10.1145/2647868.2654997. 133

Youngjae Yu, Jongseok Kim, and Gunhee Kim. A joint sequence fusion model for video question answering and retrieval. In *ECCV*, 2018. 63, 69, 72, 77

Bowen Zhang, Hexiang Hu, and Fei Sha. Cross-modal and hierarchical modeling of video and text. In *ECCV*, pages 374–390, 2018a. 64

Lei Zhang, Shuai Wang, and Bing Liu. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2018b. 30

Zhuosheng Zhang, Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, Zuchao Li, and Hai Zhao. Neural machine translation with universal visual representation. In *International Conference on Learning Representations*, 2020a. URL https://openreview.net/forum?id=Byl8hhNYPS. 66

Ziqi Zhang, Yaya Shi, Chunfeng Yuan, Bing Li, Peijin Wang, Weiming Hu, and Zheng-Jun Zha. Object relational graph with teacher-recommended learning for video captioning. In *CVPR*, 2020b. 140, 141

Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, and Yi-Dong Shen. Dual-path convolutional image-text embedding. *CoRR*, abs/1711.05535, 2017. URL http://arxiv.org/abs/1711.05535. 14, 15, 28, 32, 37, 38, 82, 91, 92, 93, 95

Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 487–495. Curran Associates, Inc., 2014. URL http://papers.nips.cc/paper/5349-learning-deep-features-for-scene-recognition-using-places-database.pdf. 91

Luowei Zhou, Chenliang Xu, and Jason J Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI Conference on Artificial Intelligence*, 2018a. URL https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17344. 20, 21, 55

Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer. In *CVPR*, 2018b. 55, 141

Mingyang Zhou, Runxiang Cheng, Yong Jae Lee, and Zhou Yu. A visual attention grounding neural model for multimodal machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3643–3653, Brussels, Belgium, October 2018c. Association for Computational Linguistics. doi: 10.18653/v1/D18-1400. URL https://www.aclweb.org/anthology/D18-1400. 9, 118

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2242–2251, 2017. doi: 10.1109/ICCV.2017.244. URL https://doi.org/10.1109/ICCV.2017.244. 24

Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *CVPR*, 2020. 55